An Automatic Child-Directed Speech Detector for the Study of Child Language Development

Soroush Vosoughi, Deb Roy

The Media Laboratory
Massachusetts Institute of Technology, USA
soroush@mit.edu, dkroy@media.mit.edu

Abstract

In this paper, we present an automatic child-directed speech detection system to be used in the study of child language development. Child-directed speech (CDS) is speech that is directed by caregivers towards infants. It is not uncommon for corpora used in child language development studies to have a combination of CDS and non-CDS. As the size of the corpora used in these studies grow, manual annotation of CDS becomes impractical. Our automatic CDS detector addresses this issue.

The focus of this paper is to propose and evaluate different sets of features for the detection of CDS, using several off-the-shelf classifiers. First, we look at the performance of a set of acoustic features. We continue by combining these acoustic features with several linguistic and eventually contextual features. Using the full set of features, our CDS detector was able to correctly identify CDS with an accuracy of .88 and F1 score of .87 using Naive Bayes.

Index Terms: motherese, automatic, child-directed speech, infant-directed speech, adult-directed speech, prosody, language development

1. Introduction

Child-directed speech (CDS), also known as *motherese* is speech that is directed by caregivers towards infants. Numerous studies in the field of child language development have shown this speech to have exaggerated prosody and shorter utterance lengths when compared to adult-directed speech (ADS) [1, 2, 3, 4, 5]. In particular, CDS on average has been shown to have higher mean fundamental frequency (F0), higher F0 range and higher vowel durations [5]. There are a few theories as to the reasons behind these differences between CDS and ADS, one theory being that prosodic exaggeration in CDS helps attract the attention of infants [2].

Moreover, studies have shown particular properties of CDS to contribute greatly to language development [6, 7, 8]. Therefore, when analyzing large datasets that include both CDS and ADS, it is very important for child language researchers to be able to separate CDS from ADS in an efficient manner so they can analyze CDS and its contributions to language development. Given the specialized nature of the CDS classification problem, there has been few prior work on it [9, 10, 11]. The importance of identifying CDS for the study of child language development and the relatively small number of prior studies on this has led us to develop a fully automatic CDS detector using the Human Speechome corpus [12].

2. Corpus

The corpus used for this paper is from the Human Speechome Project (HSP) which was collected for the study of early child language development. The HSP corpus is high-density, longitudinal and naturalistic. The corpus consists of high fidelity audio and video recordings throughout the home of a family with a young child (recorded continuously for the first three years of the child's life) [12]. Audio was recorded using ceiling mounted boundary-layer microphones (AKG C562CM) at 16 bit resolution with a sampling rate of 48 KHz. The boundary-layer microphones allowed for most of speech throughout the house to be captured. The corpus consists of roughly 120,000 hours of audio. However, our current analysis of the HSP data is on the child's 9-24 month age range which contains about 4260 hours of 14-track audio of which about 1150 hours contain speech. The data consists of adult-directed and child-directed speech from the three primary caregivers. All child-directed speech is directed at the same infant. Two of the caregivers, the mother and the father, are native English speakers while the third caregiver, the female nanny, speaks English as a second language, though this corpus consists of only English utterances.

There are approximately 2.5 million utterances in our dataset (an utterance is on average 3 seconds long). Of the 2.5 million utterances in the corpus we analyze an evenly-sampled 5000 utterances that have been hand-transcribed using new, semi-automatic methods and for which the speaker has been automatically identified with high confidence using our automatic speaker-identification system [13]. The 5000 utterances were distributed between 4 annotators who then used an annotation tool to identify the utterances as child-directed or adult-directed. This annotation tool allowed the annotators to listen to an utterance while reading the corresponding transcription and then making a decision on whether the speech was directed at the child or at an adult. In order to measure the accuracy of the human annotations, a total of 300 utterances were randomly chosen from the 5000 utterances and were given to all the 4 annotators.

Table 1 shows the inter-annotator agreement between pairs of annotators. The average pairwise inter-annotator agreement was 0.95 which shows a high level of consistency and accuracy in the human annotated data. Of the 5000 utterances, the annotators identified 3618 as child-directed and 1382 as adult-directed. In order to have an even sample of child-directed and adult-directed utterances (which will help us later in our analysis), we randomly selected 1382 of the 3618 child-directed utterances and threw out the rest. This left us with a total of 2764 utterances (about 138 minutes of audio), evenly distributed between child-directed and adult-directed speech.

Table 1: Pairwise inter-annotator agreement for all 4 annotators.

	A2	A3	A4
A1	.95	.94	.96
A2		.96	.95
A3			.97

3. CDS Detection

CDS detection is a binary classification problem. As with most classification problems, the main factor affecting the performance of our CDS detector was the set of features that were used. In this study we created three different CDS detectors using different sets of features. The first detector uses only acoustic features. The second detector relies on both acoustic and linguistic features. Finally, the third detector uses a combination of acoustic, linguistic, and contextual features. We decided to break down our features into these three sets so that we can clearly show the contribution of each set and most importantly to allow our detectors to be used on different datasets, including those that might not have access to one or more of these feature sets (e.g. audio corpora with no transcripts).

Since this paper's focus is on identifying features that can best characterize CDS and not on novel classification techniques, we used several off-the-shelf classifiers to create our CDS detectors and test the performance of these features. In this paper we only show the results for the top three performing classifiers: Naive Bayes, k-Nearest Neighbors (kNN), and Decision Tree.

3.1. Acoustic features

As mentioned in the introduction, one of the main motivating factors for this study was previous work in the field of child language development on prosodic difference between CDS and ADS [1, 2, 3, 4, 5]. Therefore for our acoustic features we decided to use the prosodic features that these papers have identified as being exaggerated in CDS. These prosodic features are: *mean fundamental frequency (F0)*, *F0 range*, *mean intensity*, and *mean vowel duration*. Since our dataset contains speech from multiple speakers we also included the *speaker* as a feature.

Since one of the main goals of an automatic CDS detector is to improve efficiency by automating the annotation of speech, we made our feature extraction process completely automatic. Figure 2 shows the complete feature extraction pipeline used in this study. Below we explain how each of these acoustic features were extracted.

3.1.1. Feature Extraction

As mentioned previously, the speaker of an utterance is identified using an automated speaker identification system that was developed in our lab [13]. F0 and intensity contours for each utterance are extracted using the PRAAT system [14]. These contours are then used to calculate the F0 range ($F0_{max} - F0_{min}$), mean F0, and mean intensity of an utterance.

In order to calculate the mean vowel duration in an utterance, we first extracted duration for all vowel tokens in our corpus using a forced-aligner. We next converted these to normalized units for each vowel separately (via z-score), and then measured the mean standardized vowel duration for all the vowels in our utterance. For example, a high score on this measure

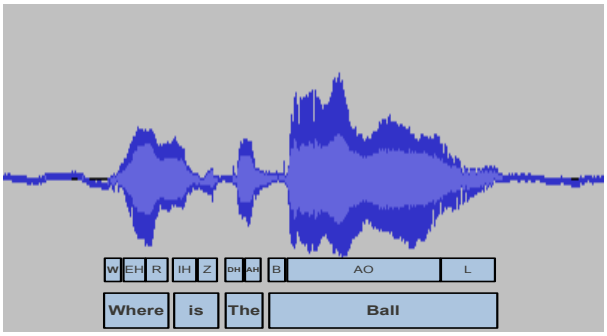

Figure 1: *A sample phoneme level alignment of an utterance generated by the HTK forced-aligner.*

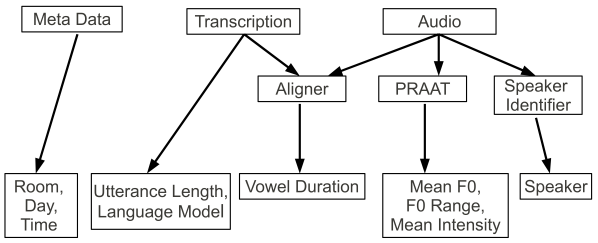

Figure 2: *Schematic of the feature extraction pipeline.*

for an utterance would reflect that the vowels that occurred in that utterance were on average often long relative to comparable vowel sounds that appeared in other utterances spoken by the same speaker. We grouped similar vowels by converting transcripts to phonemes via the CMU pronunciation dictionary [15].

The forced-aligner that was used for this task uses the Hidden Markov Model Toolkit (HTK) [16] and works as follows. Given a speech segment and its transcript, the Viterbi algorithm in HTK is applied to align the transcript to the speech segment on the phoneme level, using the CMU pronunciation dictionary [15] and an acoustic model. Since each transcript is associated with a speaker ID label that is generated automatically using our speaker-identification program [13], we can use speaker-specific acoustic models (which we have trained using thousands of samples from our corpus) to get more accurate alignments [17]. Figure 1 shows a sample phoneme level alignment of an utterance done by our forced-aligner.

3.1.2. Classification

We used 5 fold cross-validation to train and evaluate several classifiers using the acoustic features explained above. Table 2 shows the performance of each of the classifiers. For reference we included a majority classifier to act as the baseline. Since the dataset had equal number of positive and negative examples, the majority classifier has a .5 accuracy, the same accuracy as random chance. Naive Bayes was the best performing classifier with an accuracy of .74 and an area under ROC curve (AUC) of .8. Figure 3 shows the corresponding ROC curves.

3.2. Linguistic features

Not all audio datasets will have transcriptions but some datasets like the HSP corpus do have a great number of transcribed utter-

Table 2: Performance of the CDS detector using acoustic features.

	Accuracy	AUC	F1 score
Majority	.50	.50	.67
Decision Tree	.54	.57	.67
kNN(k=10)	.67	.75	.68
Naive Bayes	**.74**	**.80**	**.72**

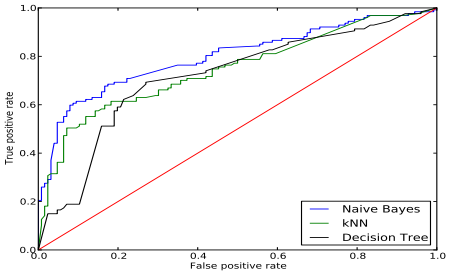

Figure 3: *ROC curves of the top three classifiers using acoustic features.*

ances. In this section we extracted and used several linguistic features from these transcripts to improve the performance of our CDS detector.

3.2.1. Feature Extraction

The first feature that we extracted from the transcripts was the number of words in an utterance, also known as the length of an utterance (*LU*). This is a good feature to use since CDS has been shown to have on average a lower *LU* when compared to ADS.

The transcripts were also used to train language models. We constructed two bigram language models for each of the three speakers, the first model was trained on CDS and the other model on ADS. A bigram language model calculates the probability of a word given the preceding word, as shown in the equation below.

$$P(w_n|w_1, w_2, w_3...w_{n-1}) = P(w_n|w_{n-1})$$

We then calculated the probability of an utterance being child-directed or adult-directed using our bigram language models, as shown in the equations below.

$$P(CDS|utt, speaker) = P(utt|bigram_{CDS,speaker})$$
$$P(ADS|utt, speaker) = P(utt|bigram_{ADS,speaker})$$

These two probability values and length of utterance (*LU*) were used as the linguistic features in our CDS detector, which combined with the acoustic features gave us a total of 8 features: *mean fundamental frequency (F0)*, *F0 range*, *mean intensity*, *mean vowel duration*, *speaker*, *LU*, *P(CDS|utt,speaker)*, and *P(ADS|utt,speaker)*.

3.2.2. Classification

As with the previous section, we used 5 fold cross-validation to train and evaluate several classifiers using the acoustic and linguistic features. Table 3 shows the performance of each of the

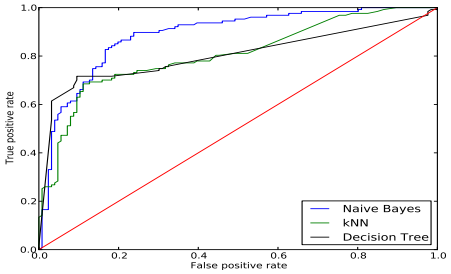

Figure 4: *ROC curves of the top three classifiers using acoustic and linguistic features.*

classifiers. Naive Bayes was still the best performing classifier with an accuracy of .82 and an AUC of .88. Figure 4 shows the ROC curves for these classifiers. The results clearly show that using linguistic features in combination with the acoustic features greatly improved the performance of our CDS detector, achieving an impressive .82 accuracy rate.

Table 3: Performance of the CDS detector using a combination of acoustic and linguistic features.

	Accuracy	AUC	F1 score
Majority	.50	.50	.67
kNN(k=10)	.74	.80	.75
Decision Tree	.78	.79	.77
Naive Bayes	**.82**	**.88**	**.82**

3.3. Contextual features

Finally, the unique nature of our dataset granted us access to certain contextual features. Specifically, since the data is recorded in the same household, each utterance is tagged with location (*room*), time of day (*ToD*), and date (which we used to find the day of the week or *DoW*). These features capture certain contextual aspects of CDS. For instance, there are certain times of the day when the caregivers are more likely to be interacting with the child. Similarly, there might be certain days of the week and rooms in which the caregivers are more likely to be interacting with the child.

Since all the speech in our corpus is tagged with this information when recorded, no further processing is needed to extract these features. Combining these features with the acoustic and linguistic features we get a total of 11 features: *mean fundamental frequency (F0)*, *F0 range*, *mean intensity*, *mean vowel duration*, *speaker*, *LU*, *P(CDS|utt,speaker)*, *P(ADS|utt,speaker)*, *ToD*, *DoW*, and *room*.

We again used 5 fold cross-validation to train and evaluate our classifiers using these 11 features. Table 4 shows the performance of our classifiers. The best classifier, Naive Bayes, achieved an accuracy of .88 with an AUC of .92. Figure 5 shows the ROC curves for these classifiers. We saw significant improvement in the performance of our CDS classifiers when using contextual features in conjunction with the other features.

Table 5 summarizes the performance of our best classifier (Naive Bayes) using the different feature sets explained in this paper. It is evident from the results that with the addition of each

Table 4: Performance of the CDS detector using a combination of acoustic, linguistic, and contextual features.

	Accuracy	AUC	F1 score
Majority	.50	.50	.67
kNN(k=10)	.78	.87	.80
Decision Tree	.86	.88	.85
Naive Bayes	**.88**	**.92**	**.87**

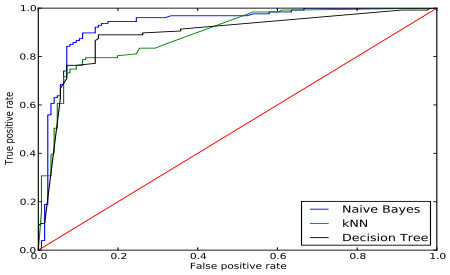

Figure 5: *ROC curves of the top three classifiers using acoustic, linguistic, and contextual features.*

of these feature sets, the performance of our CDS classifier was improved.

Table 5: Performance of the best performing classifier (Naive Bayes), using different sets of features (Acs: Acoustic, Ling: Linguistic, Ctx: Contextual).

Features	Accuracy	AUC	F1 score
Acs	.74	.80	.73
Acs + Ling	.82	.88	.82
Acs + Ling + Ctx	.88	.92	.87

4. Discussion

In this paper, we have shown how to design a fully automatic CDS detection system. We started by training and evaluating a CDS classifier using acoustic features. We then introduced linguistic and eventually contextual features to our classifier. The performance of our CDS classifier improved significantly with the addition of each of these feature sets. Our final CDS detector, trained and evaluated using the HSP corpus, was able to correctly identify CDS with an accuracy of .88 and an F1 score of .87. This is especially impressive given the relatively small size of our training dataset and the fact that the HSP corpus contains natural, noisy and unstructured speech recorded from open air microphones. An accuracy of .88 is high enough for our CDS detector to be used on the HSP corpus. Almost all previous work on CDS classification has relied exclusively on acoustic features [9, 10, 11], our CDS classifier is different from those as it uses a combination of acoustic, linguistic, and contextual features, allowing it to be used on a variety of datasets.

Our CDS detector enables child language researchers to automatically separate CDS and non-CDS. Our system will be especially useful when dealing with large datasets like the HSP corpus which currently has more than 2.5 million utterances. Overall, we see our CDS detector as an important and useful tool for research in the field of child language development.

5. Acknowledgements

Thanks to all the annotators. Thanks to Brandon Roy and Matt Miller for their work on the speaker identification system and finally thanks to Karina Lundahl for her administrative support.

6. References

[1] O. Garnica, "Some prosodic and paralinguistic features of speech to young children," *Talking to children: Language input and acquisition*, pp. 63–88, 1977.

[2] A. Fernald and T. Simon, "Expanded intonation contours in mothers' speech to newborns." *Developmental Psychology*, vol. 20, no. 1, pp. 104–113, 1984.

[3] 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. 14–20, 1984.

[4] D. D. Albin and C. H. Echols, "Characteristics of stressed and word-final syllables in infant-directed speech: Implications for word-level segmentation." *Infant Behavior and Development*, vol. 19, pp. 401–418, 1996.

[5] S. Vosoughi and D. Roy, "A longitudinal study of prosodic exaggeration in child-directed speech," in *Proceedings of Interspeech*, Shanghai, China, 2012.

[6] J. Huttenlocher, W. Haight, A. Bryk, M. Seltzer, and T. Lyons, "Early vocabulary growth: Relation to language input and gender," *Developmental Psychology*, vol. 27, 1991.

[7] C. Echols and E. Newport, "The role of stress and position in determining first words," *Language acquisition*, vol. 2, 1992.

[8] S. Vosoughi, B. C. Roy, M. C. Frank, and D. Roy, "Contributions of prosodic and distributional features of caregivers' speech in early word learning," in *Proceedings of the 32nd Annual Cognitive Science Conference*, 2010.

[9] A. Mahdhaoui, M. Chetouani, and C. Zong, "Motherese detection based on segmental and supra-segmental features," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, dec. 2008, pp. 1 –4.

[10] A. Robinson-Mosher and B. Scassellati, "Prosody recognition in male infant-directed speech," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, sept.-2 oct. 2004, pp. 2209 – 2214 vol.3.

[11] T. Inoue, R. Nakagawa, M. Kondou, T. Koga, and K. Shinohara, "Discrimination between mothers infant- and adult-directed speech using hidden markov models," *Neuroscience Research*, vol. 70, no. 1, pp. 62 – 70, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168010211000307

[12] D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, M. Levit, and P. Gorniak, "The Human Speechome Project," in *Proceedings of the 28th Annual Cognitive Science Conference*. Mahwah, NJ: Lawrence Earlbaum, 2006, pp. 2059–2064.

[13] B. C. Roy and D. Roy, "Fast transcription of unstructured audio recordings," in *Proceedings of Interspeech*, Brighton, England, 2009.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.01)," http://www.praat.org/, 2009.

[15] H. Weide., "The CMU Pronunciation Dictionary, release 0.6," Carnegie Mellon University, 1998.

[16] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Cambridge University Engineering Dept, 2001.

[17] S. Vosoughi, "Interactions of caregiver speech and early word learning in the speechome corpus: Computational explorations," MIT M.Sc. thesis, 2010.