Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio–Visual Information

Serdar Yildirim and Shrikanth Narayanan, *Senior Member, IEEE*

***Abstract*—The presence of disfluencies in spontaneous speech, while poses a challenge for robust automatic recognition, also offers means for gaining additional insights into understanding a speaker's communicative and cognitive state. This paper analyzes disfluencies in children's spontaneous speech, in the context of spoken dialog based computer game play, and addresses the automatic detection of disfluency boundaries. Although several approaches have been proposed to detect disfluencies in speech, relatively little work has been done to utilize visual information to improve the performance and robustness of the disfluency detection system. This paper describes the use of visual information along with prosodic and language information to detect the presence of disfluencies in a child's computer-directed speech and shows how these information sources can be integrated to increase the overall information available for disfluency detection. The experimental results on our children's multimodal dialog corpus indicate that disfluency detection accuracy of over 80% can be obtained by utilizing audio–visual information. Specifically, results showed that the addition of visual information to prosody and language features yield relative improvements in disfluency detection error rates of 3.6% and 6.3%, respectively, for information fusion at the feature level and decision level.**

***Index Terms*—Disfluency detection, feature selection, information fusion, spontaneous children speech, spoken language processing.**

I. Introduction

PROCESSING and understanding spontaneous speech is one of the key challenges in creating conversational interfaces. Spontaneous speech is different from read or rehearsed speech partly because of the disfluencies produced by a speaker during spoken language production. It is hence important to address the presence of disfluency in speech processing. Detection of disfluencies plays a crucial role in two aspects of speech interface design. The first is its effect on robust language understanding since a speaker's communicative intent needs to be inferred from the fluent part of the speech. The second is in dialog design because of the communication function of disfluencies in a spoken interaction, particularly in turn taking and dialog coordination. While there is considerable research on how to detect and handle disfluencies produced by an adult speaker, relatively little work has been done for that of children. The speech and speaking styles of children are different from those of adults, both in speech only and multimodal communication scenarios. This is true for speech patterns in children's spoken interactions with computers. This paper provides details on the disfluencies occurring in spontaneous speech of children. Furthermore, it addresses the problem of automatically detecting the disfluency boundaries by means of audio–visual information. The proposed signal-based methods are simple in the sense that they can operate on features extracted from the audio and video signals without requiring explicit speech or gesture recognition schemes. Note that there has been work on disfluency and miscue detection in the context of reading by children [1]–[3]. The problem, and the engineering solutions there are however different from the one addressed in this paper which focuses on disfluency detection in spontaneous speech of children where the text corresponding to the speech is not available *a priori*.

Several studies have been carried out on detecting disfluencies in spontaneous speech by either analyzing the underlying word sequence that uttered or by using acoustic–prosodic characteristics of speech. In a seminal thesis [4], Shriberg laid a comprehensive foundation highlighting the patterns in a variety of disfluency features in spontaneous speech, and their significance for speech processing applications. Stolcke and Shriberg [5] introduced a hidden event language model in which disfluency events were taken into account to estimate the probability of a word sequence. They modified the standard (n-gram) language model in a way that the probability estimate of a word that follows a hidden event was conditioned on the fluent version of a word sequence. Even though they observed no significant reduction in automatic speech recognition (ASR) word error rate by employing the hidden event language model compared to that of a standard language model, as they suggested, hidden event language model can be used to predict disfluencies in a given text. Shriberg *et al.* [6] investigated the use of prosodic cues to detect different types of disfluencies. In their approach,

Manuscript received July 13, 2007; revised October 15, 2008. Current version published December 11, 2008. This work was supported in part by the National Science Foundation (NSF) through the Integrated Media Systems Center, an NSF Engineering Research Center, Cooperative Agreement under Contract EEC-9529152, a CAREER award, and the Department of the Army under Contract DAAD 19-99-D-0046. The work of S. Yildirim was supported in part by the National Science Foundation, a USC Zumberge Interdisciplinary Research award, and a USC Annenberg Communications Critical Pathway Fellowship. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Helen Meng.

S. Yildirim was with the Department of Electrical Engineering and IMSC, University of Southern California, Los Angeles, CA 90089 USA. He is now with the Department of Computer Engineering, Mustafa Kemal University, Antakya, Hatay 31040, Turkey (e-mail: serdar@alumni.usc.edu).

S. Narayanan is with the Department of Electrical Engineering and IMSC, University of Southern California, Los Angeles, CA 90089 USA (e-mail: shri@sipi.usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2006728

they first extracted a variety of prosodic features in a local region around an inter-word boundary, with time marks obtained from forced alignments of speech to true transcription. Then, they constructed a prosody model using a classification and regression trees (CART)-style decision tree. They reported accuracies of 89.7%, 77.5%, 75.5%, and 74.0%, respectively, for the detection of filled pauses, repetitions, repairs, and false starts. In later work [7], prosodic information and hidden event-based language model were combined to detect different disfluency types and sentence boundaries based on recognized words using different combination techniques such as model interpolation, independent model combination and joint modeling. Their results show significant improvements over individual information sources. Recently, Liu *et al.* [8] proposed a novel hidden-event part-of-speech (POS) language model to capture syntactically generalized patterns in addition to acoustic–prosodic features and word-based hidden event language model in their interruption point detection system. They achieved the best results when information sources were combined. They also proposed a repetition pattern language model to capture the less frequent repetitions that can occur in a speaker's utterance.

It should be noted that all these previous efforts primarily targeted adult speech. It is well known that acoustic characteristics of children speech are different than that of adult speech in many ways. Age-dependent changes in acoustic speech parameters [9], and greater variability in their values, have serious implications for the design of robust speech applications. Results have shown that in such cases the word error rates are typically two to five times worse for children speech than for adult speech [10]–[14]. Given that, a disfluency detection system that relies on explicit segmentation information from ASR will suffer from a high word error rate and boundary alignment errors. It has been also shown that older children change their speaking style when they interact with the spoken interface, for instance they produce much less disfluencies compared to that of produced in an interpersonal spoken interaction [15]. However, this is not the case for younger (including preliterate) children, our targeted population, where their observed disfluency rates in a computer interaction setting is comparable to that produced by older children in interpersonal communication setting (see Section III for details). Higher disfluency rates can aggravate the problems of ASR or forced-alignment-based speech segmentation. One ASR-free approach to address this issue, is to use pitch breaks to mark candidate disfluency regions. Recently, Wang and Narayanan [16] proposed a multipass linear fold algorithm for sentence boundary detection using only pitch breaks and their durations. They obtained an overall error rate of 25% with a false alarm rate of 17.95% on a subset of Switchboard corpus. However, their algorithm only relies on pitch breaks and pitch durations. In this paper, we not only use pitch related features but also other prosodic features related to speech signal energy and duration.

It is also well-known that interpersonal communication is multimodal, predominantly characterized by speech combined with gestures such as hand and head movements, facial expressions, and gaze. The relation between speech and gesture in human communication has been examined extensively. It has been shown that gestures such as hand movements [17], [18] and rigid head movements and facial expressions [19], [20] are correlated with prosody and discourse structure. Analysis of the frequency of gestures, specifically hand gestures, during fluent and hesitant phases in speech for different temporal resolutions showed that speech and gestures are co-expressive and complementary [21]. Cassell *et al.* showed that posture is correlated with discourse state [22]. It has also been shown that modification gesture patterns have a high correlation with content replacement speech repairs [23]. Kettebekov *et al.* [24] presented a framework in which visual and speech signal features were combined to improve continuous gesture recognition. They obtained about 16% relative improvement over using only visual signal. Even though a number of studies have used information from speech channel to improve gesture recognition, relatively few studies have attempted to include information from the visual channel for improving the recognition of discourse characteristics of speech such as disfluencies and sentence boundaries. Recently, Chen *et al.* [25] showed improvements by including gestural information to prosody and language information in detecting sentence boundaries. They followed the direct modeling approach where explicit gesture features were calculated from measurements obtained from video signal. In this paper, we investigate the usefulness of visual cues for detecting disfluencies and explore if these cues present complementary information to acoustic cues in improving the performance of disfluency detection system. However, similar to the case of robust speech recognition, the problem of robust recognition and modeling of dynamic gestures is still not completely solved and faces a number of technological challenges. Hence, for the purpose of this work, we adopted a simple approach that does require explicit direct modeling of gestures. While there are variety of approaches to exploit visual features, for this work, we chose to use an optical flow technique in which the motion properties are directly estimated based on motion intensity changes between video frames.

Many audio–visual processing studies have shown that, in general, combining evidence from different information sources leads to performance improvements in classification. The combination can be performed at different levels. In this paper, both feature and decision level integration techniques are evaluated. For classification, three different learning techniques, Bayes-normal, k-nearest neighborhood (k-NN), and logistic model trees (LMT) [26], are considered. It is also known that features that are irrelevant for the task can hurt the classifier performance. To address this issue, sequential backward feature selection (SBFS) and principal component analysis (PCA) were performed to choose the relevant features. The algorithms were evaluated on a corpus of child–machine spontaneous spoken dialogs.

The rest of this paper is organized as follows. Section II describes the multimodal database that was constructed and used in this study. Section III analyzes the disfluency patterns in young children's speech. The approach used for the disfluency detection problem, including how each information source is modeled, and methods for feature selection are given in Section IV. Details on how to combine knowledge sources are provided in Section V. Experimental results and conclusions are given in Sections VI and VII, respectively.

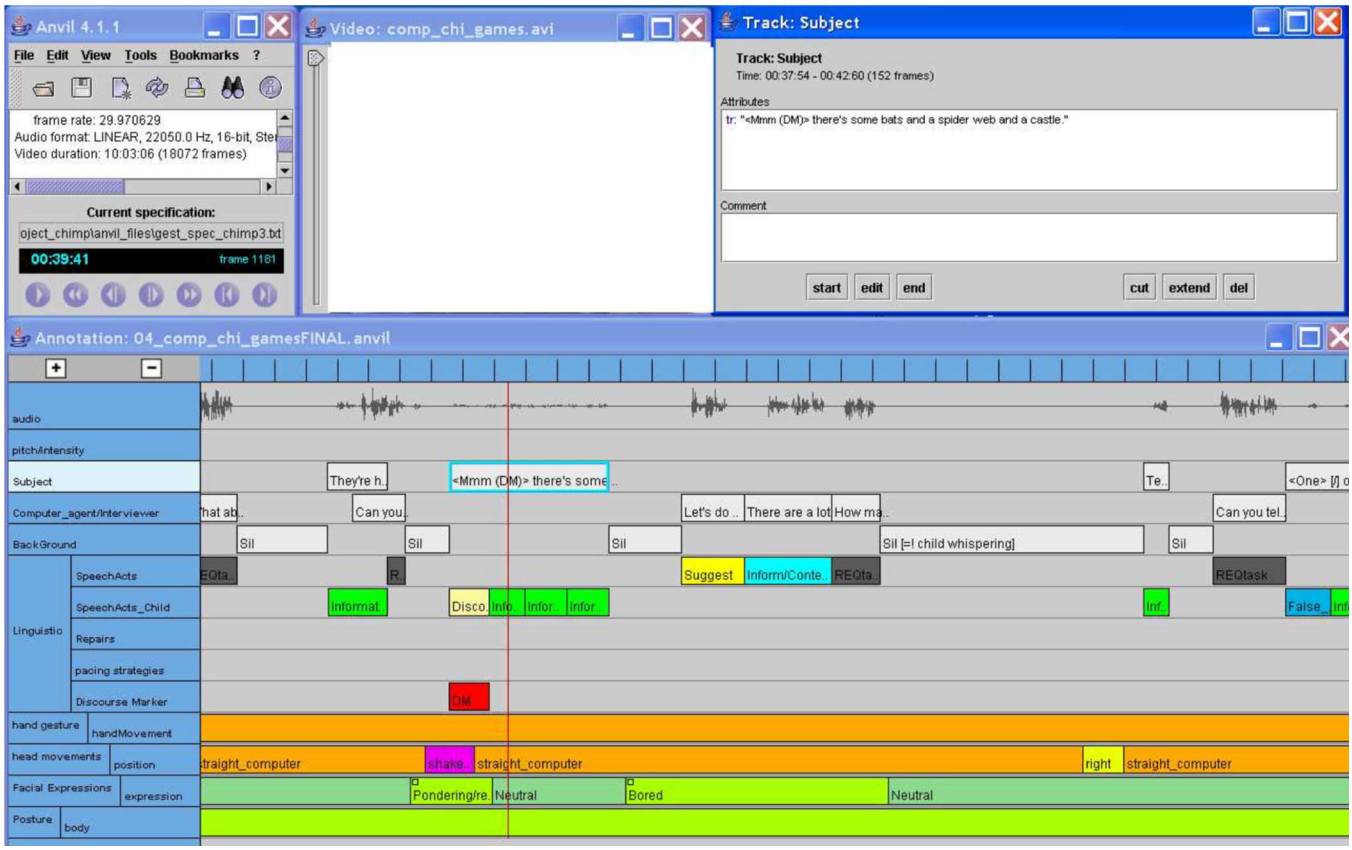

Fig. 1. Screen shot of the annotation board. The inset video screen is intentionally left blank (to preserve subject identity).

II. Multimodal Children's Database

An important requirement for designing and evaluating most data-driven information processing systems is the availability of transcribed and annotated data. For analyzing and modeling conversational behavior of children in multimodal communication scenarios, we constructed a database that consisted of audio and video recordings from ten children aged 4–6 years while interacting with a spoken dialog agent performing a series of age-appropriate cognitive tasks such as pattern recognition, sorting, and finding category membership. Further details are provided below.

A. Data Collection

For data collection, a Wizard of Oz (WoZ) tool was designed, which enabled scheduling and replay of audio/video events and allowed for a hidden human control the actions of a computer agent; the content itself was a combination of recorded and synthesized speech and audio prompts, and graphics. This procedure enabled careful control of experimental parameters including speech interpretation and machine response patterns, thus allowing for the collection of a variety of child–computer interactions. The agent plug-in was done using the CSLU Toolkit [27].

High-quality audio recordings of the child's voice were collected using a directional desktop microphone at 44.1 kHz. The interactions were also simultaneously recorded using two Sony TRV330 digital cameras, one focused only on the child from the front and the other capturing the child and the computer screen from the side.

B. Transcription and Annotation

The transcription and annotation of the audio–video data were carried out in several stages. First, audio data from each session were transcribed, using a modified version of the Childes format [9], by a native speaker of English, and further double-checked by a second native speaker of English. Next, the transcriptions were imported, utterance-by-utterance, into the Praat tool [28] to allow for aligning of the transcribed material with their acoustic counterpart. The output of this process was further imported into a multilayer annotation tool to encode additional verbal and nonverbal information and to synchronize with the visual information. A multilayer annotation board was constructed using the Anvil multimodal annotation tool [29]. Along with the speech transcription and acoustic information (e.g., pitch and intensity contours), discourse information, such as repairs (e.g., repetitions, clarifications, corrections), speech acts (e.g., opening, providing information, acknowledging) and pacing strategies (e.g., topic termination, anticipated response, topic shift) as well as gestural information, such as hand/head movements (e.g., pointing, nodding, yes/no type head shakes), body postures and facial expressions were encoded in a synchronized multilayer manner. Example screenshot from our annotation is shown in Fig. 1. Analysis of the relations between some of the discourse characteristics and gestures can be found in our previous work [30], [31].

TABLE I
DISFLUENCY TYPES CONSIDERED IN THIS STUDY AND CORRESPONDING EXAMPLES FROM THE DATABASE

Disfluency Type	Example
Repetition	One is hiding in a baseball helmet and one's hiding <in> [/] in a box.
Repair	I don't know <what> [//] who used the ladder.
False start	<I wanted to> [/-] When is Margie coming?
Filled pauses	He's <uhm> (DM) waving his hand.

TABLE II
DISFLUENCY RATE PER 100 WORDS AND THE AVERAGE NUMBER OF WORDS PER UTTERANCE FOR EACH SUBJECT. ALU: AVERAGE LENGTH OF UTTERANCE

Subject Age	Disfluency Rate (%)	ALU (a)	ALU (b)
4	4.72	5.8	5.5
4	5.50	4.7	4.4
4	10.50	7.5	6.7
5	3.26	5.2	5.0
5	8.70	4.8	4.3
5	13.60	8.6	7.3
6	3.28	5.6	5.4
6	4.60	5.7	5.4
6	7.03	6.4	5.9
6	13.10	6.0	5.8

III. DISFLUENCY IN YOUNG CHILDREN'S SPEECH

Disfluencies produced by a speaker can be associated with the cognitive and communicative state of the speaker even though they might be interpreted as speech errors. Studies show that disfluency rates in longer utterances are higher than that of shorter utterances [32], [33] suggesting that the increase in cognitive load causes the speaker to produce more disfluencies. Shriberg [33] also showed that disfluencies are more likely to occur at the beginning of an utterance rather than the later part indicating that disfluencies can also be associated with difficulty in speech planning. There is also a wealth of linguistic and psycholinguistic literature on the role of disfluencies as a communication function in a spoken interaction. Disfluencies, especially fillers such as *uh*, *uhm*, and *oh* help interlocutors to coordinate interactions and manage turn taking [33]. Given the link between disfluencies and the cognitive and communicative state of the speaker as an interlocutor in a spoken interaction, recognizing disfluencies can help the computer to predict the cognitive and communicative states of a user so that the interaction between a system and an user will be more adaptive and natural. For instance, the application can provide additional information to a child user when his/her speech becomes more disfluent since an increase in disfluency rate might be an indication of increasing cognitive load.

Despite the potential importance of natural computer interfaces for younger (preliterate) population of 4–6 years, little is known about their discourse characteristics, both in speech only and multimodal communication scenarios. Disfluency types considered in this study that targets this particular population segment, with corresponding examples drawn from the database are given in Table I. These are the most frequent disfluency types observed in this particular database.

Table II shows the disfluency rates per 100 words and the average length of utterances (average number of words per utterance (a) with disfluencies in the utterance (b) when disfluencies removed from the utterances) for each subject age 4- to 6-years old. As can be seen from the table there, is a great variability between subjects in terms of disfluency rates. Average disfluency rate of 7.43% with a standard deviation of 3.8% is observed. It has been previously reported that the older children (ages 6–10 years) produced three times higher disfluencies during interpersonal communication than while interacting with computers (6.71% versus 2.32%) [15]. Oviatt and Adams argued that one possible reason behind this difference is that 6- to 10-year-old children can distinguish an animated character from real human partners: They are possibly aware of the limitations of the computer system and change their speaking styles accordingly. When we compare our results with the those from this older age group, the average disfluency rate produced by our younger age group (4- to 6-year olds) while interacting with spoken computer interface is comparable to that produced by the older age group in an interpersonal communication setting. If we follow the argument in [15], one could speculate that these younger children fail to take into account the informational needs of the computer agent and they do not change their speaking style. One implication of this result is that interfaces addressed to young children should be programmed to adjust to the child's interactional styles. We also found a statistically significant positive correlation between the disfluency rate and the average length of utterance ($r = 0.74, p = 0.014$) in agreement with previously published results for adults and older children. This result indicates that the more the children try to construct longer utterances, the more they produce disfluencies.

IV. AUTOMATIC DISFLUENCY DETECTION

The disfluency detection problem is generally defined as automatically finding the location of the points at which disfluent speech becomes fluent in an utterance. For example, consider the following utterance,

$$\text{I don't know } \mathbf{what} * \text{ who uses the ladder.}$$

Here, the speaker first uttered the word "what" and then corrected it with "who"; the boundary between these two words is the disfluent boundary while the other interword boundaries are fluent boundaries. Note that classifying the type of disfluency is not considered in this work.

A block diagram of our approach is shown in Fig. 2. The first step is to find a set of candidate disfluent boundaries, and our method is based on pitch breaks. In most previous approaches, the interword boundary information is obtained first, either directly from the forced alignments of speech to true transcription or from the output of an automatic speech recognizer. Following that, a variety of prosodic features are extracted for each interword boundary in a local region around the estimated boundary point. However, in most cases the true (expected) transcription may not be available; furthermore, word alignment errors are unavoidable with current ASR techniques. In the case of children's speech, this could be exacerbated due to the higher word error rate compared to that of adult speech. To circumvent this

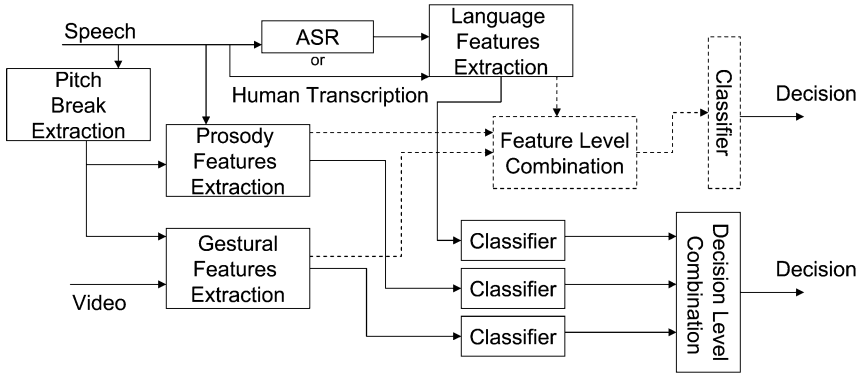

Fig. 2. Approach to disfluency detection. Dashed lines represent feature level integration, whereas solid lines represent decision level combination.

TABLE III
EVALUATION OF PITCH BREAKS IN TERMS OF HOW WELL THEY MATCH WITH THE TRUE WORD BOUNDARIES

Δt	Precision	Recall
100 ms	0.77	0.69

problem, in this study, we consider a method that does not rely on ASR based alignment, but one that uses pitch values instead. Specifically, we consider each pitch break as a candidate disfluent boundary and extract a variety of prosodic and gestural features in the local neighborhood around these pitch breaks (the region spanning from the beginning of a pitch contour that precedes the pitch break to the end of a pitch contour that follows the pitch break).

We employ the following procedure to see how well pitch breaks and transcribed word boundaries match. We consider pitch breaks match with the transcribed word boundaries if there is a pitch break within the Δt vicinity of t_i (the time interval $t_i - \Delta t < t_i < t_i + \Delta t$), where t_i is the ending time point of true word boundary. Δt was empirically chosen to be 100 ms (a randomly selected held out dataset was used to determine this value). 98% of the disfluent boundaries match with the pitch breaks at this value. The overall performance of the matching procedure is evaluated using Recall (r) and Precision (p) metrics and is shown in Table III.

$$r = \frac{\text{Number of Correctly Found Boundaries}}{\text{Total Number of Boundaries}}$$
$$p = \frac{\text{Number of Correctly Found Boundaries}}{\text{Number of Hypothesized Boundaries}}.$$

In the next step, prosody, lexical and gestural information are extracted and modeled based on the time marks obtained from the pitch-break processing for disfluency detection. The specific details are given below.

A. Prosodic Model

Acoustic prosodic features have been used, and shown to be useful, in a variety of speech processing tasks such as sentence boundary and topic segmentation, dialog act classification, syllable stress detection, expressive speech processing, and disfluency detection. To model prosody for the task of disfluency detection, we extracted a number of prosodic features around each detected pitch break. The Praat tool was used to obtain pitch, energy, and voicing measurements. The resulting raw pitch tracks were smoothed using a three-point median filter. Details of the features used are summarized below.

- *Duration:* pitch break duration, and durations of preceding and following voiced regions.
- *Pitch:* Mean, median, maximum, minimum, standard deviation, and range of the preceding and the following voiced regions, the difference in pitch features across a pitch break, linear regression coefficients of preceding and the following voiced regions. Features related to the first and second derivatives are also included.
- *Energy:* Mean, median, maximum, minimum, standard deviation, and range of preceding and the following voiced regions, the difference in energy features across a pitch break, linear regression coefficients of preceding and the following voiced regions. Features related to the first and second derivatives are also included.

Pitch and energy features were normalized with each individual speaker's baseline. A total of 83 features were obtained to model prosody.

B. Lexical Model

Our lexical model relies on annotated transcriptions. We trained a trigram hidden-event language model [5] on transcriptions that had disfluency tags inserted in the sentence, for example

I don't know what ⟨DT⟩ who uses the ladder.

Then, hidden event posteriors were estimated using the SRI Language Modeling Toolkit's (SRILM) *hidden-ngram* function [34]. We used hidden event posteriors as lexical features.

C. Gestural Model

Recent findings indicate that gestural patterns might be used as an additional source to identify speech disfluencies produced by a speaker in a spoken interaction [23]. Also in our previous work [31], co-analysis of gestural use and speech repairs revealed that children exhibited specific patterns in their interactions. Younger children were more likely to interact multimodally than just verbally than older children; furthermore, children were less likely to produce disfluencies, especially filled pauses, when they were interacting multimodally [31].

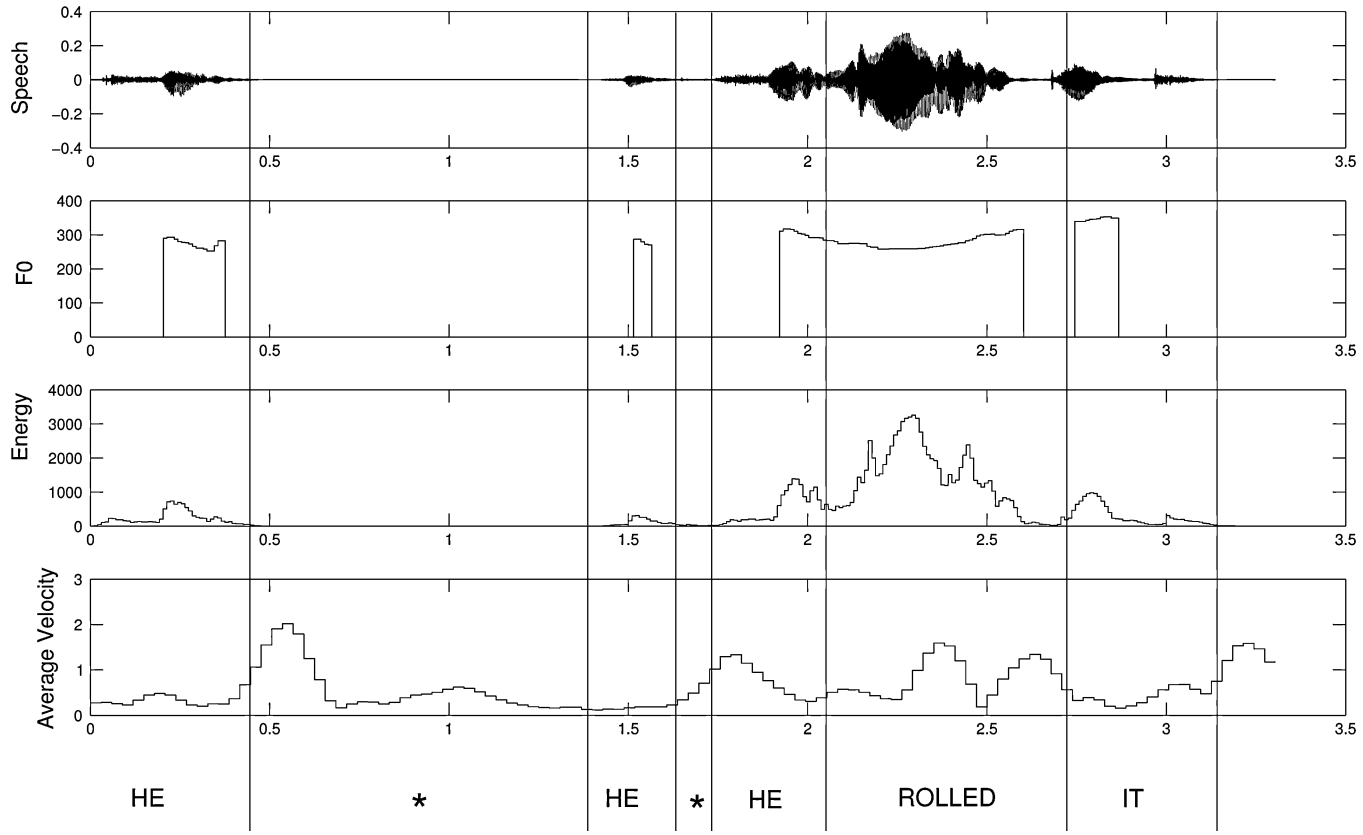

Fig. 3. Example speech waveform, pitch contour, rms energy, and average speed trajectory obtained from the corresponding video sequence for an utterance "He he he rolled it." spoken by a 4–year-old male child. Vertical lines represent word boundaries and "$*$" represents a disfluent boundary.

In summary, the availability of gesture information can be potentially advantageously used in detecting disfluency.

To model gestures for the task of disfluency detection, first gestural measurements should be obtained from the visual channel. It is well known that highly accurate gestural measurements can be obtained by employing wearable tracking devices. However, this method, which is intrusive, is not suitable for many scenarios, including the targeted children age group. Instead, we obtained the gestural measurements directly from the video sequence by using the optical flow technique in which the motion properties were estimated based on motion intensity changes between frames [35], [36]. Our goal here is to capture the information related to the child's movements (i.e., gesture formations); explicit recognition of hand and body gestures are considerably more complex, and are not the object of this work. Note that the child–computer interaction application that we consider is especially suited for this simple analysis, since the child sitting in front of the computer screen was the main object in view of the camera and the motion intensity changes that occurs between video frames were generally attributable to the gestural formations.

We use a gradient-based method to estimate the image velocity $\mathbf{v}$ where $\mathbf{v} = (v_x, v_y)^T$. This method assumes that intensity is conversed and derives the optical flow constraint equation, i.e.,

$$\nabla I \cdot \mathbf{v} + \frac{\partial I}{\partial t} = 0 \tag{1}$$

where $\nabla I = (I_x, I_y)^T$ is the spatial intensity gradient and $(\partial I)/(\partial t)$ is the partial temporal derivative of image intensity I. Image velocity can be estimated by solving (1). For each frame, the average velocity was calculated by averaging the corresponding velocities of all pixels. Average overall velocity of each frame was also calculated [37].

To model the gesture, we calculated a variety of gestural features around each pitch break from the vertical, horizontal, and average speed trajectories including the mean, median, maximum, minimum, standard deviation, and range of the preceding and the following voiced regions, and the difference between these features across a pitch break. The mean, median, maximum, minimum, standard deviation, and range of the speed trajectories of the pitch break region were also added to feature vector. A total of 62 features were obtained to model the gesture information.

An illustrative example showing the speech waveform, pitch contour, rms energy, and the average speed trajectory obtained from the corresponding video sequence for an utterance is shown in Fig. 3. The figure corresponds to a sequence where a child points to a region on the computer screen several times (each peak seen in the average speed plot corresponds to one pointing gesture) while speaking. It can be seen that both disfluency points falls into pitch break regions.

D. Feature Selection and Reduction

Some of the prosodic and visual features summarized above may be irrelevant for disfluency detection, and therefore they

can hurt the classifier performance. Also, when information sources are combined at feature level, due to the increase in the input dimension, the newly combined feature set may not necessarily result in an improved performance (curse of dimensionality). In this study, to address these issues, we applied sequential backward feature selection (SBFS) and principal component analysis (PCA) techniques.

1) Sequential Backward Feature Selection: SBFS is a feature selection method that reduces dimensionality by selecting a subset of existing features that maximize the performance of the classifier. SBFS starts with the complete set of features and removes the worst feature according to some criterion one by one until the number of remaining features reaches the preset number. Nearest neighbor classifier recognition rate was used as the feature selection criterion. Feature selection procedure can be summarized as follows.

- Compute criterion function for all k features.
- Remove each feature one at a time, compute criterion function for all subsets with $k-1$ features, and delete the worst feature (which maximizes the accuracy by removing it from the feature set)
- Remove each feature one at a time from the remaining $k-1$ features, and delete the worst feature to form a subset with $k-2$ features.
- Continue until predefined number of features are left.

2) Principal Component Analysis: An alternative to feature selection that reduces dimensionality by selecting subsets of features is feature reduction that reduces dimensionality by combining features. Here, the goal is to find linear transformation that project the high-dimensional data onto a lower dimensional space. PCA is one of the classical approach for this purpose [38].

In PCA, given z is the whole feature set, first the covariance matrix $K^{k\times k}$ is constructed, and its eigenvalues and eigenvectors are calculated. Then the new matrix $A^{n\times n}$ is constructed by using only the n largest eigenvalues and the corresponding eigenvectors. The new feature set y of size $n(n<k)$ is obtained by means of the operation given by

$$y = A^T(z-\mu) \tag{2}$$

where μ is the mean vector of z.

V. Combining Information Sources

Evidence from different information sources can be combined at different levels, from raw data to decision level. The goal here is to improve the performance and robustness of the classification system by using complementary and redundant information from different information sources. In this paper, we investigate the combination problem at both feature level and decision level. In feature level integration, a new high-dimensional feature vector is constructed by augmenting the different feature sets from the various information sources and a single classifier is trained using the new feature set. In decision-level integration, separate classifiers are used with different information sources and the final decision is made by combining results from individual classifiers using some rule. The specific details are given below.

A. Feature Level Integration

Let $X_i = [x_i1,\ldots,x_ik_i]$ denote the feature set from information source i, where $i=1,\ldots,N$, denote the various information sources, and k_i is the feature vector dimension of information source i. The new high-dimensional vector $Y=[X_1,\ldots,X_N]$, where $Y\in R^{\sum_i k_i}$ is obtained by simply concatenating the feature vectors from the different information sources. A single classifier is built using the new feature set to estimate $P(C_j\,|Y)$, where C_j is the class label (where $j=1,2,\ldots,K$ is the number of the class) and Y is the combined information. However, due to increase in dimensionality, the desired improvement in classification accuracy may not be achieved. Feature selection algorithms can be applied to reduce the dimensionality. Another problem in feature level integration is that feature sets from all information sources are needed to make the final decision. The absence of information from one or more sources, possibly due to communication break down, can have detrimental effects on the overall classification system. Parallel combination (decision level) of classifiers trained separately for different information sources may be better suitable when information from different domains needs to be combined.

B. Decision Level Integration

As pointed above, one of the important advantages of the decision level integration over feature level is that a decision can be made even in the absence of one or more information sources, since decision from other sources can be used to achieve the overall decision. At the decision level, different classifiers are used for each information source and the posterior probabilities generated from these classifiers are combined using some rule. Here, we used simple averaging to combine posterior probabilities from each classifier because it achieves good performance in spite of its less support by probabilistic interpretation [39]. Similar to feature level, the goal here is also to estimate $P(C_j|Y)$, but this time using posterior probabilities generated by each classifier. Let $P(C_k|X_i)$ be the posterior estimates from each classifier, then the rule can be formulated as follows:

$$P(C_k|Y) = \frac{1}{N}\sum_i P(C_k|X_i). \tag{3}$$

The final decision is made by

$$P(C_j|Y) = \arg\max_k P(C_k|Y). \tag{4}$$

VI. Experimental Results

A. Methodology

We considered each pitch break as a candidate disfluency boundary and extracted a number of prosodic and gestural features around each detected pitch break, as detailed in Section IV. Also, to evaluate the role of lexical features which

TABLE IV
CLASSIFICATION ACCURACY, IN PERCENT, FOR PROSODY AND VISUAL INFORMATION WITH DIFFERENT FEATURE SELECTION/REDUCTION APPROACHES (FOR PROSODY, $k = 11$ IN k-NEAREST NEIGHBOR, AND $k = 13$ FOR VISUAL INFORMATION). PROS: PROSODY

	Base		SBFS		PCA (80%)		PCA (90%)		PCA (95%)	
	Pros	Visual	Pros	Visual	Pros	Visual	Pros	Visual	Pros	Visual
k-NN	65.6	58.5	69.1	62.4	65.2	59.6	66.1	61.2	66.3	61.8
BayesN	73.2	66.0	76.1	67.8	71.4	58.5	72.9	63.9	73.5	65.9
LMT	72.9	66.9	-	-	-	-	-	-	-	-

TABLE V
PERFORMANCE OF CLASSIFIERS IN TERMS OF PRECISION, RECALL , AND F-MEASURE FOR EACH CLASS. RESULTS GIVEN FOR K-NN AND BAYES NORMAL CLASSIFIERS ARE BASED ON SBFS FEATURE SETS

		Prosody			Visual		
		recall	*precision*	*F-measure*	*recall*	*precision*	*F-measure*
k-NN	Fluent	67.9	69.4	68.6	54.7	64.3	59.1
	Disfluent	70.1	68.6	69.3	64.6	60.6	64.8
BayesN	Fluent	78.5	74.9	76.6	73.5	66.0	69.5
	Disfluent	73.6	77.4	75.5	62.2	70.1	65.9
LMT	Fluent	76.7	71.4	74.0	66.8	66.9	66.7
	Disfluent	69.2	74.8	71.9	67.0	66.9	66.9

have been widely used in disfluency modeling studies, we extracted the lexical features using annotated transcriptions (see Section IV-B for details). It is difficult to preselect the best classifier for a classification problem on a particular data set. Thus, we tested three different supervised classifiers: k-nearest neighbor (k-NN), Bayes-normal (linear discriminant classifier), and logistic model trees (LMT) to obtain the boundary class labels (disfluent or fluent). The k-NN classifies a test object according to most frequent class labels amongst the k-nearest neighbors. The second classifier we investigated, the Bayes-normal classifier, assumes each class has a Gaussian probability density with a common covariance matrix. The third classifier, LMT, builds classification trees with logistic regression trees at the leaves [26].

Since there is a large skew in our class sizes (the number of disfluent boundaries are relatively fewer compared to that of fluent boundaries in the data, 300 and 2384, respectively), for the purpose of experiments, we randomly divided the fluent class data into 8 disjoint sets to equalize the prior class sizes. Each fluent data set was then combined with the disfluent class data separately resulting in eight datasets. Each data set was then evaluated by tenfold cross validation. The final evaluation metrics were then calculated by averaging results from each data set. The results are presented and compared in terms of recall (r_j) where r_j is a ratio of the number of instances that are truly classified as class C_j ($C_1 =$ fluent and $C_2 =$ disfluent classes) to the number of instances belongs to class C_j in the data, precision (p_j), where p_j refers to a ratio of the number of the truly classified instances to the total number of instances classified to class C_j by the classifier, and F-measure (F_j) for each class where

$$F_j = \frac{2p_j r_j}{p_j + r_j} * 100\% \tag{5}$$

as well as the overall accuracy (Acc) of the classifiers where Acc is the ratio of the total number of instances correctly classified (instances correctly classified as fluent class plus instances correctly classified as disfluent class) to the total number of instances.

B. Feature Selection/Reduction Results

Table IV compares the performance of the different classification methods, and also shows the effects of the sequential backward feature selection (SBSF) and PCA-based feature reduction on the classifiers for both prosody and visual information sources. The number of neighborhoods for the k-NN classifiers was set to 11 for prosody data, and 13 for visual data. Those numbers were estimated by leave-one-out cross-validation on the whole database for each information source. For PCA, we tested the dimensions explaining 80%, 90%, and 95% of the sum of eigenvalues. We obtained best results when we took the dimension explaining 95% of the sum of eigenvalues. Results show that feature selection by SBFS yields relative improvements in classification accuracy of 3.96% and 2.72% for Bayes-normal classifier for prosodic and visual information, respectively, compared with the results using full feature set. For k-NN classifier, relative improvements are 5.33% and 6.67% for prosodic and visual information, respectively. Even though no significant improvement is achieved by reducing the feature dimension with PCA for both information sources over baseline, PCA showed comparable performance with the base feature sets in the reduced feature dimensions. As can be observed from Table IV, feature sets obtained by applying SBFS with Bayes-normal classifier had the best performance in terms of classification accuracy for both audio and visual information sources.

Table V shows the detailed performance of classifiers in terms of precision, recall, and F-measure for each class. For k-NN and Bayes normal, results were based on the SBFS feature sets. We reduced the dimension of feature vectors from 83 to 30 for the prosody information, and from 62 to 45 for the visual information source, respectively. We also applied SBFS one more time when the prosodic and visual feature vectors were combined at the feature level to reduce the final feature vector dimension. The combined feature set dimension was reduced from 75 to 40. These new feature set dimensions were obtained by evaluating the effect of dimension reduction on recognition accuracy using held out data.

TABLE VI
FIFTEEN BEST FEATURES SELECTED BY SBFS METHOD. SUPERSCRIPTS $(^-)$ AND $(^+)$ REFER PRECEEDING AND FOLLOWING VOICED REGIONS, RESPECTIVELY, _first AND _second REFER FIRST DERIVATIVES AND SECOND DERIVATIVES, RESPECTIVELY

Acoustic Features	Visual Features
$pitch\_break\_dur$	s_y_std
v^+_dur	s^-_mean
$F0^-\_reg\_coeff$	s_y_range
$F0^-\_mean - F0^+\_mean$	s_x_range
$F0^-\_max - F0^+\_max$	s_x_std
e^-_range	s_range
$e^+\_reg\_coeff$	$s_x^+\_median$
$F0^-\_median$	$s_x^-\_range$
$F0^+\_reg\_coeff$	s_std
$F0^+\_second\_max$	s_median
$e^-\_range - e^+\_range$	$s_x^+\_mean$
$F0^+\_first\_range$	s^-_median
$F0^-\_std$	s^+_mean
$e^+\_second\_std$	$s_y^+\_mean$
$F0^+\_second\_std$	$s_y^+\_std$

Table VI lists selected prosodic and visual features for disfluency detection in decreasing order of importance by SBFS method. According to feature selection, pitch break duration (pitch_break_dur), duration of following voiced region (v_dur^+), F0 regression coefficient of preceding voiced region $(F0^-\_\text{reg\_coeff})$, difference of preceding and following voiced region F0 mean values $(F0^-\_\text{mean} - F0^+\_\text{mean})$, difference of preceding and following voiced region F0 maximum values $(F0^-\_\text{max} - F0^+\_\text{max})$ are the first five most determinant prosodic features for disfluency detection. Gestural features extracted from pitch break region such as vertical speed standard deviation (s_y_std), vertical speed range (s_y_range), horizontal speed range (s_x_range), and horizontal speed standard deviation (s_x_std) are included in the five-best gestural features.

C. Information Combination Results

Table VII shows the performance results for the feature and decision level information techniques with different classifiers in terms of the overall classification accuracy. The classification performance based on language information is also given in this table. Overall, the best performances were achieved when all information sources were fused at decision level. The best overall accuracy of 82.1% was obtained when the Bayes-normal classifier was used with each of the three information sources and results are combined at decision level. Also, the combination of all three - prosody, lexical, and gestural—knowledge sources gave better results than compared to using either each information source alone or in any pairwise combination. It is worth noting that prosody/lexical combination leads to a larger improvement in the detection accuracy than prosody/visual combination. One possible reason is that the correlation between visual and prosodic features is higher than that between lexical and prosodic features at disfluency boundaries. It is expected that when two knowledge sources are correlated, classifiers trained using these cues may give similar decisions. In order to assess the similarity between two classifiers, we calculated the Q statistics [40]. It is defined for two classifiers, x, y as

$$Q_{x,y} = \frac{N^{00}N^{11} - N^{01}N^{10}}{N^{00}N^{11} + N^{01}N^{10}} \tag{6}$$

where N^{00} is the number of times both classifiers are wrong, N^{11} the number of times both classifiers are correct, and N^{01} is the number of times when the first is correct and second classifier is wrong, N^{10} is the number of times when the first classifier is wrong and the second classifier is correct. The Q statistic takes values between $[-1, 1]$ and closer the value to 0, the more dissimilar the classifiers are. $Q_{x,y} = 0$ represents total dissimilarity. In Table VIII, the pairwise Q statistics are given for the information sources considered in this study. As can be observed, Q statistics between classifiers based on prosody and visual cues are higher than the other pairwise information sources considered.

TABLE VII
PERFORMANCE COMPARISONS OF DECISION-LEVEL AND FEATURE-LEVEL INFORMATION COMBINATION TECHNIQUES WITH DIFFERENT CLASSIFIERS

		BayesN	k-NN	LMT
	Prosody Only	76.1	69.1	72.9
	Language Only	74.8	71.1	74.6
	Visual Only	67.8	62.2	66.9
Feature Level	Language+Visual	76.4	75.3	76.6
	Prosody+Visual	76.4	72.4	74.4
	Prosody+Language	79.8	77.2	79.8
	Prosody+Language+Visual	80.5	77.6	81.1
Decision Level	Language+Visual	76.7	74.8	76.0
	Prosody+Visual	77.5	69.5	74.5
	Prosody+Language	80.9	77.2	80.2
	Prosody+Language+Visual	82.1	78.0	81.1

TABLE VIII
PAIRWISE Q STATISTICS BETWEEN CLASSIFIERS TRAINED WITH DIFFERENT KNOWLEDGE SOURCES

Q(Prosody,Visual)	0.5176
Q(Prosody,Language)	0.2531
Q(Language,Visual)	0.2701

VII. CONCLUSION

With recent advances in multimedia technologies, there has been increasing interest in the design of multimodal interfaces for applications such as interactive games, automated tutoring, and instructional materials for children. However, building such interfaces is a challenging task and there are several issues that need to be addressed. In this paper, we explored the detection of disfluencies present in the child's speech using data obtained from children ages between 4 and 6 while interacting with computer agent by means of audio–visual information. Knowledge of meta-linguistic events such as disfluency not only can enable robust speech processing but also provide valuable insights into the cognitive state of the human.

Our analysis results showed that younger (preliterate) children produced disfluency on an average rate of 7.4% when they interacted with our spoken interface. This result shows that the disfluency rates for young children are high when compared to the previously published rates produced by older children while

communicating with an automated spoken interface. This result points out the added importance of detecting disfluencies in the context of designing computer interfaces for younger children.

In this paper, we also investigated the problem of disfluency detection using multimodal sources. We proposed a method in which pitch breaks were used as candidate disfluency regions. In addition to acoustic and lexical information, we proposed the use of visual information for the task of disfluency detection. In our previous work, we showed that children exhibited specific patterns in their interactions in terms of gestural use and speech repairs; for instance, they were less likely to produce disfluencies when they were interacting multimodally [31]. Thus, gestural information can be a potential information source for the automatic recognition of disfluency boundaries. We used a simple way for indirectly capturing gestural (motion) activity by the use of an optical flow technique instead of requiring an explicit gesture recognition system.

There are several open issues that need to be further explored in the future. The lexical model in this work relied on annotated transcriptions. Given the small size of the corpus considered, building well optimized ASR system for this data was difficult. We hope to incorporate ASR-derived lexical features into our system as we collect and transcribe more data. For visual information, we only considered a simple set of motion related features extracted based on image intensity using optical flow technique without explicit tracking of specific parts of a body. However, it has been shown that gestures such as those related to hand and head movements are highly correlated with speech. In the future, we plan to track such specific gestures from video sequence and investigate if the disfluency detection system can be improved further.

REFERENCES

[1] S. Banerjee, J. E. Beck, and J. Mostow, "Evaluating the effect of predicting oral reading miscues," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 3165–3168.

[2] K. Lee, A. Hagen, N. Romanyshyn, S. Martin, and B. Pellom, "Analysis and detection of reading miscues for interactive literacy tutors," in *Proc. 20th Int. Conf. Comput. Linguist. (Coling)*, Geneva, Switzerland, Aug. 2004, Article 1524.

[3] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young childrens speech for the purpose of assessment," in *Proc. InterSpeech ICSLP*, Antwerp, Belgium, Aug. 2007, pp. 206–209.

[4] E. E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph.D. dissertation, Univ. of California, Berkley, 1994.

[5] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. ICASSP*, Atlanta, GA, 1996, vol. 1, pp. 405–408.

[6] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. Eurospeech*, 1997, pp. 2383–2386.

[7] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Proc. ICSLP*, 1998, no. 5, pp. 2247–2250.

[8] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge source," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 957–960.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1455–1468, Mar. 1999.

[10] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 65–78, Feb. 2002.

[11] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and elderly," in *Proc. ICASSP*, 1996, pp. 349–352.

[12] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. ICASSP*, 1998, pp. 433–436.

[13] A. Hagen, B. Pellom, and R. Cole, "Childrens speech recognition with application to interactive books and tutors," in *Proc. IEEE ASRU Workshop*, St. Thomas, Virgin Islands, Dec. 2003.

[14] Q. Li and M. J. Russell, "An analysis of the causes of increased error rated in children's speech recognition," in *Proc. ICSLP*, Denver, CO, 2002, pp. 2337–2340.

[15] S. L. Oviatt and B. Adams, , J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds., "Designing and evaluating conversational interfaces with animated characters," in *Embodied Conversational Agents.*. Cambridge, MA: MIT Press, 2000, pp. 319–343.

[16] D. Wang and S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *Proc. ICASSP*, May 2004, vol. 1, pp. 525–528.

[17] D. McNeill, F. Quek, K.-E. McCullough, S. D. N. Furuyama, R. Bryll, X.-F. Ma, and R. Ansari, "Catchments, prosody, and discourse," *Gesture*, vol. 1, no. 1, pp. 9–33, 2001.

[18] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari, "Multimodal human discourse: Gesture and speech," in *ACM Trans. Comput.-Human Interaction*, 2002, vol. 9, no. 3, pp. 171–193.

[19] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 15, no. 8, pp. 2331–2347, Nov. 2007.

[20] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 15, no. 3, pp. 1075–1086, Mar. 2007.

[21] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. McCullough, and R. Bryll, "Analysis of speech and gesture frequency during fluent and hesitant phases in speech," in *Proc. 6th Multi-Conf. Syst., Cybern., Inf. (SCI 2002)*, Orlando, FL, Jul. 14–18, 2002.

[22] J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich, "Non-verbal cues for discourse structure," in *Proc. 39th Meeting Assoc. Comput. Linguist.*, Toulouse, France, Jul. 2001, pp. 106–115.

[23] L. Chen, M. Harper, and F. Quek, "Gesture patterns during speech repairs," in *Proc. ICSLP*, Denver, CO, 2002, pp. 629–632.

[24] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody based audiovisual coanalysis for coverbal gesture recognition," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 234–242, Apr. 2005.

[25] L. Chen, Y. Liu, M. Harper, and E. Shriberg, "Multimodal model integration for sentence unit detection," in *Proc. ICMI*, State College, PA, 2004, pp. 121–128.

[26] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Mach. Learn. J.*, vol. 59, no. 1–2, pp. 161–205, 2005.

[27] J. Ma, J. Yan, and R. Cole, "Cu animate tools for enabling conversations with animated characters," in *Proc. ICSLP*, 2002, vol. 1, pp. 197–200.

[28] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," Jul. 20, 2005 [Online]. Available: http://www.praat.org, (version 4.3.19) [computer program], 2005, retrieved from

[29] M. Kipp, "Anvil—A generic annotation tool for multimodal dialogue," in *Proc. Eurospeech*, 2001, pp. 1367–1370.

[30] S. Montanari, S. Yildirim, S. Khurana, M. Landes, L. Lawyer, E. Andersen, and S. Narayanan, "Analyzing the interplay between spoken language and gestural cues in conversational child–machine interactions in pre/early literate age group," in *Proc. InStil*, Jul. 2004, paper ID 047.

[31] S. Montanari, S. Yildirim, E. Andersen, and S. Narayanan, "Reference marking in childrens computer-directed speech: An integrated analysis of discourse and gesture," in *Proc. ICSLP*, Oct. 2004, pp. 1841–1844.

[32] S. Oviatt, "Predicting spoken disfluecies during human-computer interaction," *Comput. Speech Lang.*, vol. 9, pp. 19–35, 1995.

[33] E. Shriberg, "Disfluencies in switchboard," in *Proc. ICSLP*, 1996, pp. 11–14.

[34] A. Stolcke, "Srilm—An extensible language modeling toolkit," in *Proc. ICSLP*, 2002, vol. 2, pp. 901–904.

[35] J. L. Barren, D. J. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vision*, vol. 12, pp. 43–77, 1994.

[36] D. J. Fleet and K. Langley, "Recursive filters for optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 1, pp. 61–67, Jan. 1995.

[37] R. A. Peters, C. W. G. Clifford, and C. S. Evans, "Measuring the structure of dynamic visual signals," *Animal Behaviour*, vol. 64, pp. 131–146, 2002.

[38] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[39] D. Tax, M. van Breukelen, R. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying," *Pattern Recognition*, vol. 33, pp. 1475–1485, 2000.

[40] L. Kuncheva and C. Whitaker, "Measure of diversity in classifier esembles," *Mach. Learn.*, vol. 51, pp. 181–207, 2003.

Serdar Yildirim received the B.S. degree in electrical and electronics engineering from Cukurova University, Adana, Turkey, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from University of Southern California (USC), Los Angeles, in 2000 and 2006, respectively.

Currently, he is an Assistant Professor of Computer Engineering, at Mustafa Kemal University, Antakya, Turkey. His general research interests include speech recognition, spoken language understanding, spontaneous speech processing, disfluency detection, and emotion recognition in speech.

Shrikanth Narayanan (S'88–M'95–SM'02) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 19952000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 300 papers and has 14 granted/pending U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP02, ICASSP'05 MMSP'06, and MMSP'07. He is Editor for the *Computer Speech and Language Journal* (2007–present) and an Associate Editor for the *IEEE Signal Processing Magazine* and the IEEE TRANSACTIONS ON MULTIMEDIA. He was also an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004). He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication Committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.