Analysis of Linguistic Disfluencies in Bilingual Children's Discourse

Sheena Christabel Pravin, M. Palanivelan

***Abstract*—**Speech disfluencies are common in spontaneous speech. The primary purpose of this study was to distinguish linguistic disfluencies from stuttering disfluencies in bilingual Tamil–English (TE) speaking children. The secondary purpose was to determine whether their disfluencies are mediated by native language dominance and/or on an early onset of developmental stuttering at childhood. A detailed study was carried out to identify the prosodic and acoustic features that uniquely represent the disfluent regions of speech. This paper focuses on statistical modeling of repetitions, prolongations, pauses and interjections in the speech corpus encompassing bilingual spontaneous utterances from school going children – English and Tamil. Two classifiers including Hidden Markov Models (HMM) and the Multilayer Perceptron (MLP), which is a class of feed-forward artificial neural network, were compared in the classification of disfluencies. The results of the classifiers document the patterns of disfluency in spontaneous speech samples of school-aged children to distinguish between Children Who Stutter (CWS) and Children with Language Impairment CLI). The ability of the models in classifying the disfluencies was measured in terms of F-measure, Recall, and Precision.

***Keywords*—**Bilingual, children who stutter, children with language impairment, Hidden Markov Models, multi-layer perceptron, linguistic disfluencies, stuttering disfluencies.

Open Science Index, Cognitive and Language Sciences Vol:12, No:6, 2018 waset.org/Publication/10009121

I. Introduction

SPEECH disfluencies are distinguished as being either linguistic disfluencies, which pre-dominantly contain repetitions, interjections and revisions, or stuttering disfluencies, which include part of the utterance or syllable repetitions, and vowel prolongations. In [1], the authors state that all children demonstrate linguistic disfluencies; however, CWS are more disfluent and exhibit more stuttering disfluencies than the CLI. A child who produces 3% or more stuttering-type disfluencies is considered as one who stutters [3].

Syntactic complexity and discourse complexity influence the fluent production of speech [2]. According to Adams' "Demands and Capacities" model, when complexity increases, the demands for the production of fluent speech may exceed the individual's abilities; therefore, fluency may be compromised [2]. Linguistic disfluencies increase with linguistic complexity. If linguistic disfluencies in speech are produced frequently, linguistic disfluencies may indicate difficulties with utterance formulation or word finding [4]. Some research has also shown that the frequency of linguistic disfluencies and stuttering disfluencies increase in narrative contexts as opposed to conversational contexts for all children [5], [6].

Veiga et al. analyzed the acoustic characteristics of filled pauses vs. segmental prolongations in a corpus of Portuguese broadcast news, using prosodic and spectral features to discriminate between both categories [7]. A university lectures corpus subset was used in [8] with conclusions that the best features to identify whether an element should be rated as fluent or disfluent are: prosodic phrasing, contour shape, and presence/ absence of silent pauses. In [9], the authors analyze the prosodic behavior of the different regions of a disfluency sequence, pointing out prosodic contrast strategy (pitch and energy increases) between the reparandum and the repair. Some data evidenced that age might influence a number of disfluencies; its amount increases along the speaker's age. Thus, linguistic disfluencies should be categorically distinguished from stuttering disfluencies for early diagnosis of developmental stuttering and for suitable therapeutic intervention. Sequel of studies have highlighted that children with language and learning difficulties tend to use more linguistic disfluencies than do their peers [11]. To summarize, a great number of linguistic disfluencies might be a symptom of atypical language acquisition; on the other hand, production of linguistic disfluencies might be treated as a natural component of non-prepared, spontaneous discourse [12], [13]. In [14], the authors introduce a new model for detecting restart and repair disfluencies in spontaneous speech transcripts called Long Short-Term Memory-Noisy Channel Model (LSTM-NCM). The model uses a Noisy Channel Model (NCM) to generate n-best candidate disfluency analyses, and a LSTM language model to rescore the NCM analyses. Additional prosodic information is leveraged along with lexical features in [15]; the disfluencies are classified using a semi-Markov conditional random field that distinguishes disfluent chunks (to be deleted) from fluent chunks (everything else). By making chunk-level predictions, standard token-level features that can consider the entire reparandum and the start of the repair enable the model to easily capture the parallelism between these two parts of the utterance.

This paper is organized as follows: Section II describes the Speech corpus used to train and evaluate the HMM and MLP. Section III introduces the Prosodic and acoustic features. Section IV elaborates the classification using HMM and MLP. Section V describes the Evaluation phase of the experimentation. Section VI delineates the Results and Section

Sheena Christabel Pravin (Assistant Professor) is with the Department of ECE, Research Scholar (Anna University), Rajalakshmi Engineering College, Chennai, India (e-mail: sheena.s@rajalakshmi.edu.in).

M. Palanivelan (Professor) is with the Department of ECE, Rajalakshmi Engineering College, Chennai, India (e-mail: Palanivelan.m@rajalakshmi.edu.in).

VII presents our conclusions on the proposed methodology of distinguishing linguistic disfluency from stuttering disfluency.

II. The Speech Corpus

A. Database

The disfluent speech data corpus was built by collecting speech samples from10 school-aged children ranging in age between 8 years and 12 years, out of which, five were boys and five were girls. Boys pre-dominantly exhibited more disfluencies than girls in conversational speech. The corpus contains digitized waveforms and transcriptions of a large number of sessions in which the children were asked to talk on any of the topics pictorially presented to them as a visual stimulus. The subjects responded to queries in bilingual sentences for which manual and automated transcripts were prepared. Totally, 500 sentences were recorded.

TABLE I
Comprehensive Queries

Vacations	When I Grow Up
1. When was your last vacation?	1. What do you aspire to be when you grow up?
2. Describe your favorite vacation spot.	2. Will you help the needy?
3. With whom do you like to spend time with?	3. How will you spend your fortune?
4. What souvenirs do you buy for your friends?	4. Which is your dream country to pursue a career?
5. How long do you go on vacation?	5. What is your dream portfolio?

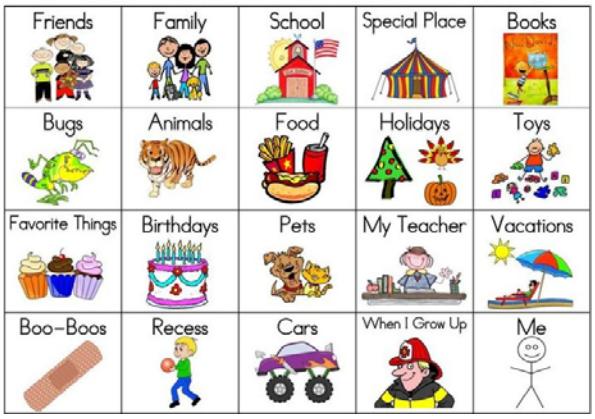


Fig. 1 Stimulus material [10]

B. Characteristics of Disfluencies

Of the 500 sentences in our corpus, 98 of the sentences contained repetitions, 85 contained false starts, 62 sentences contained hesitations, 80 sentences contained interjections, 15 sentences contained Prolongations, 50 sentences included Pauses and 25 small sentences were fluent. As well, 15% of the sentences were longer and contained multiple disfluencies.

The disfluent speech corpus was annotated using the PRAAT tool on two tiers. Tier 1 of Annotation contains the disfluency across the spoken sentences and the Tier 2 of Annotation contains the actual transcription. Table II indicates the manual assessment of disfluent regions in the spoken utterances.

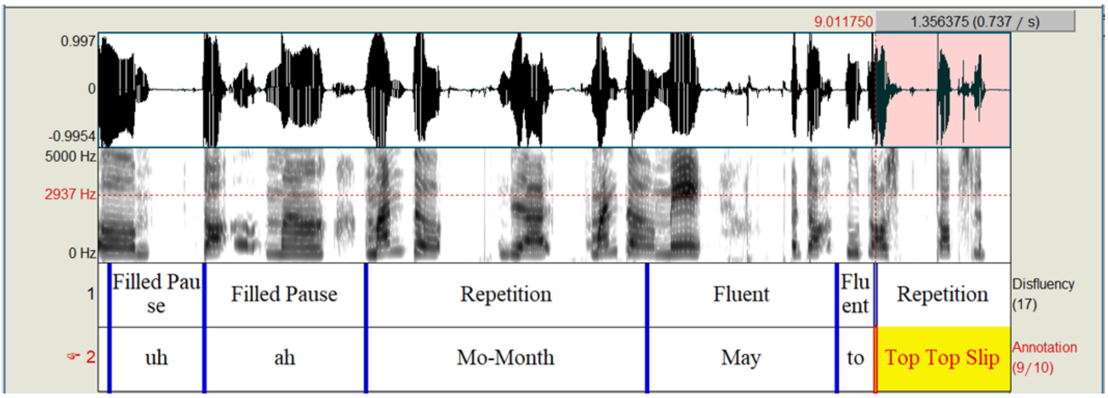


Fig. 2 An example of annotated spontaneous disfluent speech from the corpus: F_01

TABLE II
Manual Disfluency Assessment in the Spontaneous Speech

	Repetitions	Interjections	Prolongations	Pauses
M_01	15	16	7	20
M_02	20	10	8	18
F_01	12	12	9	19
F_02	5	10	7	11

III. Prosodic Features

Disfluent utterances include distinct structural regions based on acoustic and prosodic features. Prosodic features derived from the energy (energy maximum, minimum and median) of the speech signals, the estimated pitch contours, word durations, and silences before the word. Filled pauses in spontaneous speech have the stretching out trait. The vocal cord vibrates periodically and the vocal tract is observed to maintain a relatively stable contour throughout the utterance. When the speed of speaking becomes faster than the speed of preparing its content, a speaker uses filled or unfilled pauses until the next speech content resulting from the thinking process arrives at the speaking process.

IV. Classification of Disfluencies

A. Hidden Markov Model

An HMM was built for each disfluency type. The Baum-

Welch algorithm was used to find the parameters that maximize the likelihood of the observations for each HMM disfluency model. It is an iterative algorithm that converges to a locally optimal solution from the initialization values. The HMMs are uniquely described by the notation

$$\lambda = (A, B, \pi) \qquad (1)$$

where, 'A' is the state-transition probability distribution matrix, 'B' is the observation symbol probability distribution matrix, 'π' is the initial state distribution [16].

Let λ_i denote the parameter set for a disfluent class 'i'. When presented with an observation sequence, $o_1, o_2, \ldots. o_T$, the disfluency class prediction is done as:

$$\text{Predicted Disfluent Class=argmax}_i\, f(o_{1,}o_2, \ldots o_T; \lambda_i) \qquad (2)$$

The first two Formant Frequencies (F1 and F2) were chosen as features for training the HMM disfluency models. Additional prosodic features were modeled as observation likelihoods attached to the N-gram states of the HMM. The posterior probabilities were discretized before classification.

B. MLP Architecture

In a MLP, several layers composed of single units are connected to form a network of layers. Each layer in such a neural network computes a function over a vector, which is either the input to the network or the output of the previous layer. Finally, the model computes an output vector that is interpreted as p(y|x). Joined together, these individual layers compute a function that is parametrized by the connection strength between units of connected layers. These connections are usually referred to as the parameters of the neural network, or, more commonly, as its weights. A unit of a neural network layer, also sometimes referred to as neuron, computes a simple non-linear function on the weighted sum of its inputs. Even though each unit computes only a simple function, joined together, the function that is modeled by the neural network can become arbitrarily complex.

A variety of choices for the activation function f have been described in the literature. The only constraint on f is that it must be differentiable in order to enable the network to be trained using gradient decent. Although linear functions are possible, the expressive power of neural networks stems from the fact that non-linear transformations are used as activation function. Popular choices for f are the sigmoid function σ(x):

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (3)$$

In this multilabel classification task, the target vector y can therefore be considered a tuple of random variables (y1 to y_k). In the simplest form of expressing this situation, it is feasible to assume that the probability of one class is conditionally independent from the other classes. The joint distribution over labels can thus be expressed as:

$$\text{p(y|x) = p(y1, . . . , y}_\text{k}\text{|x)} = \prod p(y_k/x) \qquad (4)$$

C. Training the Neural Network

The most popular method for minimizing the cost function of a neural network is gradient descent. The idea behind gradient descent is to compute the gradient of the cost function with respect to the weights of the network in order to find the direction of steepest descent.

$$\theta \leftarrow \theta - \beta\nabla_\theta C\ (X; \theta) \qquad (5)$$

where β is the learning rate which determines the size of the step into the direction of the steepest descent, X is the training set. The cost function, C, which is to be minimized, is a function of both, the model parameters θ and the data in the training set X. An epoch is said to have passed every time the training procedure has processed the training data completely. Depending on the data set, its size and the complexity of the model along with the characteristics of the machine learning task, the data is usually processed several times over several epochs. Interestingly, conditioning the learning rate reduction on the validation set costs has shown worse performance than conditioning it on the validation set F-measure. When the model was observed to overfit the training set, the learning rate is reduced more often and converges towards zero. Table III holds the hyperparameters defined for the MLP.

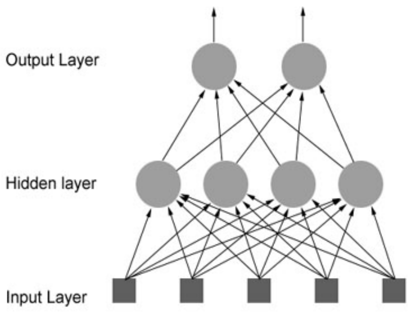


Fig. 3 Architecture of MLP

TABLE III
HYPERPARAMETERS FOR THE MLP EXPERIMENT

No. of Hidden Layers	No. of Epochs	Gradient	β	Validation Checks
70	14	0.0189	0.1	6

V. EVALUATION

Precision, which is also referred to as positive predictive value, is the fraction of relevant instances among the retrieved instances, while Recall, which is also referred as sensitivity is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.

$$Recall(R) = \frac{No.of\ True\ Positives}{No.of\ True\ Positives + No.of\ False\ Negatives} \qquad (6)$$

$$Precision(P) = \frac{No.of\ True\ Positives}{No.of\ True\ Positives + No.of\ False\ Negatives} \qquad (7)$$

$$F - measure = \frac{2\,P*R}{P+R} \tag{8}$$

The F-measure better reflects how good a classifier solves the classification task on an unbalanced data set. Since ultimately, the classifier's performance in terms of F-measure is of interest, it would be desirable to optimize for the F-measure directly.

TABLE IV
PREDICTING REPETITIONS

Classifier	Precision	Recall	F-measure
HMM	72.58	69.23	70.86
MLP	84.12	72.31	77.76

TABLE V
PREDICTING INTERJECTIONS

Classifier	Precision	Recall	F-measure
HMM	67.57	60.76	63.98
MLP	79.21	70.59	74.65

TABLE VI
PREDICTING PROLONGATIONS

Classifier	Precision	Recall	F-measure
HMM	78.57	72.5	75.41
MLP	87.21	81.03	84

TABLE VII
PREDICTING PAUSES

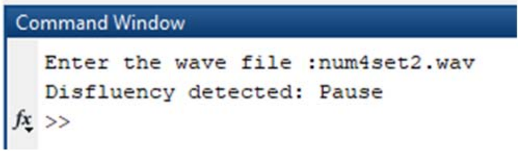
Classifier	Precision	Recall	F-measure
HMM	75.33	64.56	69.53
MLP	78.56	77.21	77.88

TABLE VIII
PREDICTING FLUENT SPEECH

Classifier	Precision	Recall	F-measure
HMM	76.23	72.32	74.22
MLP	85.47	79.42	82.33

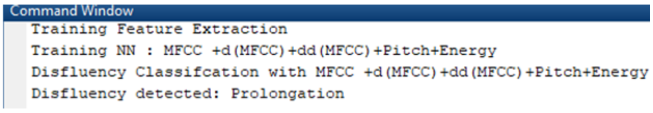
VI. RESULTS AND DISCUSSION

Filled pauses exhibit the highest F0 increase, and repetitions exhibit the highest energy. In our study we could observer that sequential filled pauses (e.g. "uh uh uh") showed successively lower starting F0 values. Fig. 4 presents the training of Repetition HMM model. Fig. 5 is the description of disfluency recognition.



```
Enter the wave file :num4set2.wav
Disfluency detected: Pause
>>
```

Fig. 4 Disfluency Classification using HMM Model



```
Training Feature Extraction
Training NN : MFCC +d(MFCC)+dd(MFCC)+Pitch+Energy
Disfluency Classifcation with MFCC +d(MFCC)+dd(MFCC)+Pitch+Energy
Disfluency detected: Prolongation
```

Fig. 5 Training and Testing Phase-MLP

TABLE IX
WORD ERROR RATE (WER) FOR AUGMENTED BASELINE PROSODIC FEATURES (PITCH CONTOUR, ENERGY, WORD DURATION) WITH ACOUSTIC FEATURES

Feature Vector	WER [%]
BASELINE	15.2
BASELINE + 13 MFCC+ ΔMFCC + ΔΔMFCC	13

Word Error Rate (WER) obtained for the Baseline Feature vector and the combination of Baseline features augmented with Prosodic features are tabulated in Table IX.

VII. CONCLUSION

The spontaneous speech of school-aged children were recorded and analyzed for distinguished assessment of linguistic disfluencies and stuttering disfluencies. Unique HMM was built for each type of disfluency and the MLP network was built of 10 input neurons indicating 10 prosodic feature inputs, 70 neurons in the hidden layer and one output. As tabulated, the MLP performed better than the HMM in classifying the disfluencies with an accuracy of 82.67%, whereas the HMM classify 70% of the disfluent utterances correctly. Child language development might be treated from the linguistic structural dimension and/ or disfluency rate dimension. Linguistic disfluencies are the basis for studying the processes of speech production in children. In this study, it is evident that fillers might not play the same functional role in the child's speech as in the adult's. The main question addressed the nature of linguistic disfluency is a manifestation of child language immaturity. But some qualitative differences between the groups were observed. For example, when faced with some problems of the utterance planning in spontaneous speech, the CWS tended to use filled hesitations, whereas the children with LD sought to make silent pauses or repetitions of part of a word. Besides that, the CWS had an inclination to produce shorter utterance to reduce the cognitive loading in utterance programming.

REFERENCES

[1] Shapiro, David Allen, "A Collaborative Journey to Fluency Freedom, 2nd Edition, Pro ed, 2011.

[2] Conture, E. G.," Stuttering: Its nature, diagnosis, and treatment" (3rd Ed.). Boston: Allyn and Bacon, 2005

[3] Adams, M., "The demands and capacities model I: theoretical elaborations. Journal of Fluency Disorders, Vol. 15, pp. 135-141, 1990.

[4] Thordardottir, E. T. & Ellis 78Weismer, S.," Content mazes and filled pauses in narrative language samples of children with specific language impairment" Brain and Cognition, 48 (2-3), 587-592, 2002.

[5] Miller, J. F., Long, S., McKinley, N., Thormann, S., Jones, M. A., & Nockerts, A., "Language Sample Analysis II: The Wisconsin guide", Madison, WI: Wisconsin Department of Public Instruction, 2005.

[6] Dehak, N.et. al., "Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification", Audio, Speech and Language Processing, September 2007, Volume 15, pp.2095-21035.

[7] A. Veiga, S. Candeias, C. Lopes and F. Perdigao, "Characterization of Hesitation using Acoustic Models, "International Congress of Phonetic Sciences – ICPhs XVII, 2011.

[8] H. Moniz, I. Trancoso and A. I. Mata, "Classification of disfluent phenomena as fluent communicative devices in specific prosodic context," in Interspeech 2009.

[9] H. Moniz, F. Batista, I. Trancosi and A. I. M. da Silva, "Prosodic context-based analysis of disfluencies," in Interspeech 2012.

[10] Guo. L., Tomblin J. B., Samelson V., "Speech disruptions in the narratives of English-speaking children with specific language

impairment, Journal of Speech, Language and Hearing Research, Vol.51(3), pp. 722-738.

[11] Akxutina T. V., ”Language Production: Neurolinguistic Syntactic Analaysis”, Moscow, 1989.

[12] Schegloff E. A, Jefferson, G.Sacks H., “Preference for self-correction in the organization of repair in conversation, Language, Vol.53(2), pp.361-381.

[13] Retrieved November 13, 2017, https://www.pinterest.com/Lcusick4288/kindergarten-writing-ideas/.

[14] Jamshid Lou, Paria & Johnson, Mark., “Disfluency Detection using a Noisy Channel Model and a Deep Neural Language Model”. 547-553. 10.18653/v1/P17-2087, 2017.

[15] Ferguson, James & Durrett, Greg & Klein, Dan., ”Disfluency Detection with a Semi-Markov Model and Prosodic Features”, 257-262. 10.3115/v1/N15-1029, 2015.

[16] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.

Sheena Christabel Pravin is currently working as a Assistant professor in the department of ECE, Rajalakshmi Engineering College Chennai. She completed her Master’s degree in Communication System and Bachelor degree in Electronics and Communication Engineering under Anna University Chennai India. Currently she is pursuing her research in the field of Linguistic and Disfluent Speech processing in Anna University Chennai India. Her areas of interest include Digital speech processing, Speech Recognition and Enhancement.

M. Palanivelan, currently working as Professor and Head at the Department of Electronics and Communication Engineering at Rajalakshmi Engineering College, Chennai, India. He has completed his Ph.D. at Anna University, Chennai at 2015. He received his Master Degree from College of Engineering (CEG), Anna university, Chennai at 2001 and Bachelor’s degree in Engineering from MS university, Tirunelveli, India at 1995. He has published several Research papers in International Journals and Conferences. He is a **Senior member IEEE** and also a life Member of ISTE and ACM. Presently guiding many research scholars under his supervision. His research interest includes Peak power problems in Multicarrier modulation systems, Multiple antenna wireless communication systems, Optical Communication, Internet of Things and Signal processing.