Classification of Speech Dysfluencies Using Speech Parameterization Techniques and Multiclass SVM

P. Mahesha[1] and D.S. Vinod[2]

[1] Department of Computer Science and Engineering
S.J. College of Engineering
Mysore, Karnataka, India
maheshsjce@yahoo.com
[2] Department of Information Science and Engineering
S.J. College of Engineering
Mysore, Karnataka, India
ds.vinod@daad-alumni.de

Abstract. Stuttering is a fluency disorder characterized by the occurrences of dysfluencies in normal flow of speech, such as repetitions, prolongations and interjection and so on. It is one of the serious problems in speech pathology. The goal of this paper is to present experimental results for the classification of three types of dysfluencies such as syllable repetition, word repetition and prolongation in stuttered speech. The three speech parameterization techniques :Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) are used as speech feature extraction methods. The performance of these parameterization techniques are compared using the results obtained by thorough experimentation. The speech samples are obtained from University College London Archive of Stuttered Speech (UCLASS). The dysfluencies are extracted from these speech samples and used for feature extraction. The multi-class Support Vector Machine (SVM) is employed for the classification of speech dysfluencies.

1 Introduction

Speech is one of the effective ways of communication between people. The basic purpose of speech is to send and receive a message in the form of language communication. Even speakers who are normally fluent experience dysfluencies due to emotional, physiological or psychological factors. Speaker is dysfluent when involuntarily repeating a word, prolonging a word, forgetting a word mid utterance, or interjecting too many "uh's" and "um's" during speech.

Stuttering is a sort of fluency disorder that sways the flow of speech. Approximately about 1% of the community is suffering from this disorder and has found to affect four times as many males as females[25,5,24,2]. The most perceptible attribute of this disorder is the production of certain types of delinquencies in

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 298–308, 2013.

Table 1. Type of Dysfluencies with Example

Type of Dysfluencies	**Example**
Repetition	
Whole word	"What-what-what are you doing"
Part word	"What t-t-t time is it?"
Prolongation	
Sound/ syllable	"I am Booooobbbby James"
Interjection (Filled pauses)	
Sound/syllable	"Um – uh -well, I had problem in morning"
Silent pauses	
Silent duration within speech considered normal	"I was going to the [pause] store"
Broken words	
A silent pause with in words	"it was won[pause]derful"
Incomplete phrase	
Grammatically in complete utterance	"I don't know how to .... let us go, guys"
Revisions	
Changed words, ideas	"There was a dog, no rat named Arthur"

the flow of speech. Dysfluencies are disturbances or breaks in the smooth flow of speech. This disorder is characterized by following major types of dysfluencies such as repetitions, prolongations, interjections, broken words etc. Examples of these dysfluencies are recorded in Table 1. Stuttering is the subject of interest to researchers from various domains like speech physiology, pathology, psychology, acoustics and signal analysis. Therefore, this subject is a multidisciplinary research field of science.

In conventional stuttering assessment method Speech Language Pathologists (SLP) classify and count the occurrence of dysfluencies manually by transcribing the recorded speech. These types of assessments are based on the knowledge and experience of speech pathologists. However, making such assessment are time consuming, subjective, inconsistent and prone to error [23,11,10,18,6]. Therefore, it would be sensible if stuttering assessment is often done through classification of dysfluencies using speech recognition technology and computational intelligence. The dysfluent speech processing is one of the areas, where research remains substantially ongoing.

In the last two decades, several studies [1,15,16,11,10,3] have been carried out on the automatic detection and classification of dysfluencies in stuttered speech by means of acoustic analysis, parametric and non-parametric feature extraction and statistical methods. Which facilitate SLPs for objective assessment of stuttering. In[1], author used Artificial Neural Network (ANN) and rough set to detect stuttering events yielding accuracy of 73.25% for ANN and about 91% for rough set. The authors of [15,16] proposed Hidden Markov Model (HMM)

based classification for automatic dysfluency detection using MFCC features and achieved 80% accuracy. In [11], automatic detection of syllable repetition was presented for objective assessment of stuttering dysfluencies based on MFCC and perceptron features. An accuracy of 83% was achieved. Subsequently in[10],same author obtained 94.35% accuracy using MFCC features and SVM classifier. Authors of [3] achieved 90% accuracy with Linear Discriminant Analysis (LDA), k- Nearest Neighbor (k-NN) and MFCC features. In[13] the same author used similar classifiers, LDA and k-NN for the recognition of repetitions and prolongations with Linear Predictive Cepstral Coefficient (LPCC) as feature extraction method and obtained the best accuracy of 89.77%. In [19]our previous work, we have developed a procedure for classification of dysfluency using MFCC feature and k-NN classifier and obtained best accuracy of 97.78% for k=5.

Investigation of the literature shows that, different feature extraction and classification algorithms have been proposed. Most of these methods concentrate on classification of syllable repetition dysfluency. In few works[5,3,9] classification of prolongation is also considered. However, stuttering is characterized by different dysfluencies as listed in Table 1. There are no attempts in literature for classifying two forms of repetition such as syllable and word repetition.

Therefore, in this work we are proposing LPC, LPCC and MFCC based speech parametrization techniques to classify three types of dysfluencies such as syllable repetition, word repetition and prolongation. The Multiclass SVM is employed for classification of dysfluencies. The comparative analysis of the these parameterization techniques are presented.

2 Database

The speech database was acquired from UCLASS[17,22]. It contains recordings of stuttered speech. This database is freely available to assist people doing research in the area of stuttered speech. We have selected 20 sound recordings for experimentation. These twenty speech files include 10 male and 10 female speakers with age ranging from 11 to 20 years. The samples were selected with intent to cover a wide range of age and stuttering rate.

3 Methodology

The classification system aims to identify the different types of dysfluencies such as syllable repetition, word repetition and prolongation. The system performs thorough analysis of speech signal by extracting features which contain characteristic information of dysfluencies. The SVM classifier is used to classify different types of dysfluencies. Our classification system has four modules : segmentation, pre-processing, features extraction and classification as shown in Figure 1.

3.1 Segmentation

The collected speech samples of UCLASS database are analyzed to identify and segment the dysfluencies manually, which is tedious but straight forward

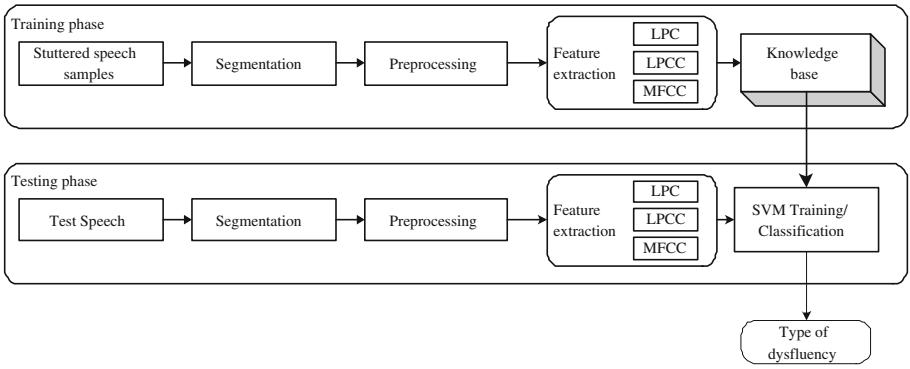

Fig. 1. Block diagram of classification

approach[[11,10]. The segmented speech syllables are subjected to feature extraction. We segmented three types of dysfluencies such as syllable repetitions, word repetitions and prolongations from speech samples.

3.2 Speech Signal Pre-processing

The speech signal pre-processing is performed to enhance the accuracy and efficiency of the feature extraction process. This phase is common for all the feature extraction methods as shown in Figure 2. This is carried to spectrally flatten the signal. A pre-emphasis filter is a simple first order high pass filter used to flatten the signal[12]. Typically the first order FIR filter is used as transfer function. The z-transform of the filter is given by

$$H(z) = 1 - \bar{a} * z^{-1}, \qquad 0.9 \leq a \leq 1.0 \tag{1}$$

The output of the pre-emphasis network $\bar{s}(n)$ is related to the input of network $s(n)$, by difference equation:

$$\bar{s}(n) = s(n) - \bar{a}s(n-1) \tag{2}$$

The output of pre-emphasized signal $\bar{s}(n)$ is divided into frames of N samples. Adjacent frames are sampled by M samples, in order to analyze each frame in the short time instead of analyzing the entire signal at once[9]. If $x_l(n)$is the l^{th}frame and there are L frames within entire speech signal, then

$$x_l(n) = s(Ml + n), \qquad n = 0, 1, \ldots, N-1 \; and \; l = 0, 1, \ldots, L-1 \tag{3}$$

The Hamming window is applied to each frame, which has the form :

$$w(n) = 0.54 - 0.46cos\left[\frac{2\pi n}{n-1}\right], \qquad 0 \leq n \leq N-1 \tag{4}$$

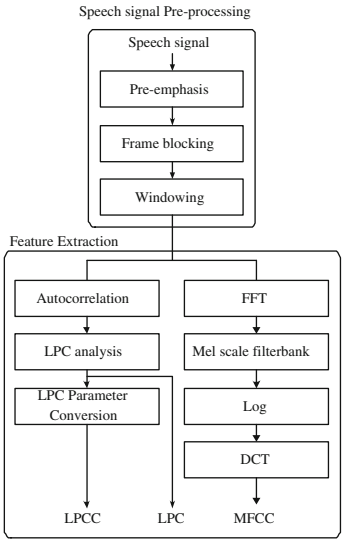

Fig. 2. Block diagram of speech signal pre-processing and feature extraction

3.3 Speech Parameterization

Speech parameterization is an important step in speech recognition systems. It is used to extract significant features from speech samples. Feature extraction is to convert an observed speech signal to some type of parametric representation for further investigation and processing. Three speech parameterization techniques were employed in this study, namely LPC, LPCC and MFCC.

LPC and LPCC. LPC is one of the most prevailing speech analysis technique. The steps involved in computation of LPC is shown in Figure 2. The LPC model is based on a mathematical approximation of the vocal tract represented by tube of a varying diameter. The key characteristic of LPC is, given the speech sample at time n, $\widehat{s}(n)$ can be predicted as linear combination of past p sample values. Where p represents order of the LPC[12].

$$\widehat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{5}$$

The prediction error $e(n)$ at any time is the difference between the actual and the estimated sample value, given by

$$e(n) = s_n - \widehat{s}(n) \tag{6}$$

$$= s_n - \sum_{i=1}^{p} a_i s(n-i) \tag{7}$$

In our work LPC with autocorrelation method is applied to each frame of windowed signal as given in [12], given by equation 8 and 9

$$r(m) = \sum_{i=1}^{N-1-m} x(n)x(n+m), \qquad m = 0, 1, \ldots p \tag{8}$$

where the autocorrelation function is symmetric, as a result the LPC equations can be stated as

$$\sum_{m=1}^{p} r(\mid m-k \mid)a_m = r(m), \qquad 1 \leq m \leq p \tag{9}$$

LPCC is Linear Prediction Coefficients (LPC) represented in the cepstrum domain [7]. These are the coefficients of the Fourier transform representation of the log magnitude spectrum. After obtaining LPC we compute Cepstral Coefficients(CC). LPCC can be derived directly from the LPC coefficients set. The recursion used is defined as follows :

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad 1 \leq m \leq p \tag{10}$$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad m > p \tag{11}$$

c_m- Cepstral coefficients, a_m- Predictor coefficients, k - $1 < k < N-1$, p - pth order.

MFCC. The MFCC is one of the popular speech parameterization technique and most commonly used feature for speech recognition. It produces a multidimensional feature vector for every frame of speech. In this study we have considered 12 MFCCs. The method is based on human hearing perceptions which cannot perceive frequencies over 1KHz. In other words, MFCC is based on known variation of the human ear's critical bandwidth with frequency[14]. The block diagram for computing MFCC is illustrated in Figure 2.

In first step, Fast Fourier Transform(FFT) is applied to pre-emphasized signal to convert each frame of N samples from time domain to frequency domain. Then, a set of triangular filters also called Mel-scale filters are used to compute a weighted sum of filter spectral components and the output of the process approximates to a Mel scale. The Mel frequency scale is linear up to 1000 Hz and logarithmic there after[20]. The Mapping of linear frequency to Mel scale is represented by the following equation (12). In final step log Mel spectrum is converted back to time domain using Discrete Cosine Transform (DCT). The outcome of conversion is called MFCCs.

$$mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{12}$$

In this study, 12 LPC, LPCC and MFCC were extracted to classify the three types of dysfluencies, namely syllable repetition, word repetition and prolongation in stuttered speech.

3.4 SVM Training and Classification

We have used SVM method for classifying three different types of dysfluencies. A SVM is a classification technique based on the statistical learning theory[21,4]. It is supervised learning technique that uses a labeled data set for training and tries to find a decision function that classifies best the training data. The purpose of the algorithm is to find a hyperplane to define decision boundaries separating between data points of different classes. It is commonly used in pattern recognition and classification problem. It gives good classification performance with limited training data compared to other classifiers. The hyper plane equation is given by

$$w^T x + b \tag{13}$$

where w is weight vector and b is bias.

Given the training labeled data set$\{x_i, y_i\}_{i=1}^{N}$with $x_i \in \mathbb{R}^d$ being the input vector and $y_i \in \{-1, +1\}$.Where $x_i$is input vector and y_i is its corresponding label[8]. SVMs map the d-dimensional input vector x from the input space to the d_h- dimensional feature space by non-linear function $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_h^d$. Hence hyperplane equation becomes

$$w^T \varphi(x) + b = 0 \tag{14}$$

with $b \in \mathbb{R}$ and w an unknown vector with the same dimension as $\varphi(x)$. The resulting optimization problem for SVM, is written as

$$\min_{w,\xi,b} \jmath_1(w, \xi) = \frac{1}{2} w^T w + c \sum_{i=1}^{n} \xi_i \tag{15}$$

such that

$$y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N \tag{16}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, N \tag{17}$$

The constrained optimization problem in equation 15, 16 and 17 is referred as the primal optimization problem. The optimization problem of SVM is usually written in dual space by introducing restriction in the minimizing function using Lagrange multipliers. The dual formulation of the problem is

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i, x_j) \tag{18}$$

subject to $\alpha_i \geq 0$ for all $i = 1, \ldots m$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$.

Thus, the hyperplane can be written in the dual optimization problem as :

$$f(x) = sgn\left[\sum_{i=1}^{m} y_i \alpha_i \left(x_i, x\right) + b\right] \tag{19}$$

Multiclass SVM. In this study multiclass SVM classifies given testing sample to one of the three classes. There are several methods for dealing with multiple classes. In the current work we use "one vs rest" method where, element that belong to a class are differentiated from the other. We calculate an optimal hyperplane that separates each class from the rest of the elements.

To get classification of N classes, a set of binary classifiers are constructed where each training separates one class from the rest. After that we combine them by doing the multiclass classification according to the maximal output before applying the *sgn* function, that takes a form

$$\arg \max_{j=1,\ldots M} g^j(x), \quad where\ g^j(x) = \sum_{i=1}^{m} y_i \alpha_i^j k(x, x_i) + b^j \tag{20}$$

and

$$f^j(x) = sgn(g^j(x)) \tag{21}$$

This has a linear complexity as for N classes we compute N hyperplanes. In our study we have 3 classes and we compute 3 hyperplanes.

4 Experimental Results

As explained in the section 2, speech samples are selected from UCLASS database. From the selected speech samples, we have created 150 speech segments of syllable repetition, word repetition and prolongation. Using these set of segments, we created a training and a testing group. The 80% of the segment is used for training and 20% for testing. The Table 2 shows the distribution of speech segments for training and testing.

To extract features form the speech samples, we have considered three speech parameterization techniques namely LPC, LPCC and MFCC. The experiment is conducted independently for each of the features by considering the same data

Table 2. The speech data

	Speech segments	Training	Testing
Syllable repetition	50	40	10
Word repetition	50	40	10
Prolongation	50	40	10
Total	150	120	30

as given in Table 2. We use SVM for the classification of dysfluencies and the total 60 speech segments are divided into 3 classes. In each experiment, we chose 80% of each class as the training set and remaining 20% as the testing data. The experiment was repeated 3 times, each time different training and testing sets were built randomly. The average accuracy of each type of dysfluencies were compared and reported in Figure 3.

Table 3 shows the classification result for syllable repetition, word repetition and prolongation with three different types of feature extraction techniques for three different set. It also shows the average accuracy of three dysfluencies for each set and overall average accuracy for LPC, LPCC and MFCC.

Table 3. The average classification accuracy of LPC, LPCC and MFCC using multiclass SVM

	LPC			**LPCC**			**MFCC**		
Dysfluencies	*Set 1*	*Set 2*	*Set 3*	*Set 1*	*Set 2*	*Set 3*	*Set 1*	*Set 2*	*Set 3*
Syllable repetition	60%	60%	80%	100%	80%	100%	100%	80%	100%
Word repetition	60%	80%	60%	80%	90%	100%	60%	80%	80%
Prolongation	80%	100%	100%	100%	100%	80%	100%	100%	100%
Accuracy	66.00%	80.00%	80.00%	93.00%	90.00%	93.00%	86.00%	86.00%	93.00%
Avg Accuracy	75.00%			92.00%			88.00%		

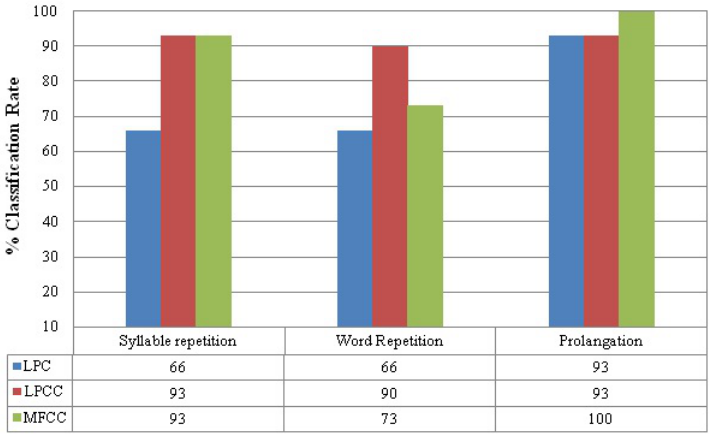

Fig. 3. Average classification of each dysfluencies for LPC, LPCC and MFCC features

5 Conclusion

In this work, effectiveness of three speech parameterization techniques such as LPC, LPCC and MFCC are investigated in categorization of syllable repetition,

word repetition and prolongation dysfluencies in stuttered speech. Multiclass SVM is used to perform classification. The recognition accuracy for parameterization techniques such as LPC, LPCC and MFCC is 75.00%, 92.00% and 88.00% respectively for all the three dysfluencies.

The experimental results demonstrate that the LPCC and SVM based system slightly outperforms because it has more discriminating capability. There is enough scope for extending the SVM with different kernel function to experiment on a larger corpus of dysfluent speech samples as part of future investigation.

References

1. Czyzewski, A., Kaczmarek, A., Kostek, B.: Intelligent processing of stuttered speech, vol. 21, pp. 143–171 (2003)
2. Bloodstein, O.: A handbook on stuttering. Singular Publishing Group,Inc., San-Diego (1995)
3. Chee, L.S., Ai, O.C., Hariharan, M., Yaacob, S.: MFCC based recognition of repetition and prolongation in stuttered speech using k-nn and lda. In: Proccedings of 2009 IEEE Student Conference on Research and Development (SCOReD), Malaysia (November 2009)
4. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press (2000)
5. Sherman, D.: Clinical and experimental use of the iowa scale of severity of stuttering. Journal of Speech and Hearing Disorders, 316–320 (1952)
6. Noth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., Wittenberg, T.: Automatic stuttering recognition using hidden markov models. Interspeech (2000)
7. Antoniol, G., Rollo, V.F., Venturi, G.: Linear predictive coding and cepstrumcoefficients for mining time variant information from software repositories. In: Proceedings of the 2005 International Workshop on Mining Software Repositories (2005)
8. Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., Suykens, J.: A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal. Chim. Acta 665, 129–145 (2010)
9. Proakis, J.G., Manolakis, D.G.: Digital signal processing. principles, algorithms and applications. MacMillan, New York
10. Ravikumar, K.M., Reddy, B., Rajagopal, R., Nagaraj, H.: Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. In: Proceedings of World Academy Science, Engineering and Technology, pp. 270–273 (2008)
11. Ravikumar, K.M., Rajagopal, R., Nagaraj, H.C.: An approach for objective assessment of stuttered speech using MFCC features. ICGST International Journal on Digital Signal Processing DSP 9, 19–24 (2009)
12. Rabiner, L., Juang, B.: Fundamentals of speech recognition. Prentice hall (1993)
13. Sin Chee, L., Chia Ai, O., Hariharan, M., Yaacob, S.: Automatic detection of prolongations and repetitions using lpcc. In: Proccedings of International Conference for Technical Postgraduates, TECHPOS (2009)
14. Lindasalwa, M., Begam, K.M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. Journal of Computing 2, 138–143 (2010)

15. Wisniewski, M., Kuniszyk-Jozkowiak, W., Smolka, E., Suszynsk, W.: Automatic detection of disorders in a continuous speech with the hidden markov models approach. In: Computer Recognition Systems 2. ASC, vol. 45, pp. 445–453. Springer, Heidelberg (2008)
16. Wisniewski, M., Kuniszyk-Jozkowiak, W., Smolka, E., Suszynski, W.: Automatic detection of prolonged fricative phonemes with the hidden markov models approach. Journal of Medical Informatics & Technologies 11 (2007)
17. Howell, P., Huckvale, M.: Facilities to assist people to research into stammered speech. Stammering Research, 130–242 (2004); an Online Journal Published by the British Stammering Association
18. Howell, P., Sackin, S., Glenn, K.: Development of a two stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: Ii. ann recognition of repetitions and prolongations with supplied word segment markers. Journal of Speech, Language, and Hearing Research 40, 1085 (1997)
19. Mahesha, P., Vinod, D.S.: Automatic classification of dysfluencies in stuttered speech using MFCC. In: Proccedings of International Conference on Computing Communication & Information Technology (ICCCIT), Chennai, India (June 2012)
20. Prahallad, K.: Speech technology: A practical introduction topic: Spectrogram, cepstrumand mel-frequency analysis. Technical report, JCarnegie Mellon University and International Institute of Information Technology, Hyderabad
21. Schoslkopf, B., Smola, A.: Learning with kernals, support vector machines. MIT Press, London (2002)
22. Devis, S., Howell, P., Batrip, J.: The UCLASS archive of stuttered speech. Journal of Speech (April 2009)
23. SAwad, S.: The application of digital speech processing to stuttering therapy. In: Proceedings of Instrumentation and Measurement Technology Conference: IEEE Sensing, Processing, Networking, pp. 1361–1367 (1997)
24. Cullinan, W.L., Prathe, E.M., Williams, D.: Comparison of procedures for scaling severity of stuttering. Journal of Speech and Hearing Research, 187–194 (1963)
25. Young, M.A.: Predicting ratings of severity of stuttering (monograph), pp. 31–54 (1961)