INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Portable Automatic *PA-TA-KA* Syllable Detection System to Derive Biomarkers for Neurological Disorders

Fei Tao[1], *Louis Daudet*[2], *Christian Poellabauer*[2], *Sandra L. Schneider*[3] *and Carlos Busso*[1]

[1] Multimodal Signal Processing (MSP) Lab, The University of Texas at Dallas, Richardson TX
[2] Mobile Computing Lab (M-LAB), University of Notre Dame, Notre Dame, IN
[3] Department of Communicative Sciences and Disorders, Saint Mary's College, Notre Dame, IN

fxt120230@utdallas.edu,ldaudet@nd.edu,cpoellab@nd.edu,sschneider@saintmarys.edu,busso@utdallas.edu

Abstract

Neurological disorders disrupt brain functions, affecting the life of many individuals. Conventional neurological disorder diagnosis methods require inconvenient and expensive devices. Several studies have identified speech biomarkers that are informative of neurological disorders, so speech-based interfaces can provide effective, convenient and affordable prescreening tools for diagnosis. We have investigated stand-alone automatic speech-based assessment tools for portable devices. Our current data collection protocol includes seven brief tests for which we have developed specialized *automatic speech recognition* (ASR) systems. The most challenging task from an ASR perspective is a popular diadochokinetic test consisting of fast repetitions of "PA-TA-KA", where subjects tend to alter, replace, insert or skip syllables. This paper presents our efforts to build a speech-based application specific for this task, where the computation is fast, efficient, and accurate on a portable device, not in the cloud. The tool recognizes the target syllables, providing phonetic alignment. This information is crucial to reliably estimate biomarkers such as the number of repetitions, insertions, mispronunciations, and temporal prosodic structure of the repetitions. We train and evaluate the application for two neurological disorders: *traumatic brain injuries* (TBIs) and Parkinson's disease. The results show low syllable error rates and high boundary detection, across populations.

Index Terms: automatic speech recognition, neurological disorders diagnosis

1. Introduction

Neurological disorders affect the nervous systems, disrupting the normal function of the brain, spine or nerves connected between them. There are more than 600 diseases related to neurological disorders, such as *Parkinson's disease* (PD), *Alzheimer's disease* (AD), stroke, *traumatic brain injuries* (TBIs), and brain tumors. Individuals with these diseases have low quality of life with discouraging prognosis, as the condition progresses affecting cognitive, communicative and motor functions. Current medical techniques normally diagnose the neurological disorders relying on expensive and intrusive devices such as *nuclear magnetic resonance* (NMR) systems, CT scan and X-ray scan. The high cost and technical skill required to operate these devices are other barriers. Currently there are no systems or devices available for personal use in early diagnosis of the disease or for monitoring the progression of the condition. Due to these reasons, many diseases caused by brain injuries such as concussions are not medically detected, and the individuals do not receive appropriate treatment. New alternative methods are needed that are effective, portable, convenient and affordable.

This work was funded by the NSF under grant IIS-1450349.

Speech production is a complex process that involves about 100 different muscles and triggers about 140000 neuromuscular events, per second [1]. The disruption caused by neurological disorders affects speech production, showing deviations in acoustic features when compared with speech from a typical healthy person [2]. These speech-based biomarkers can serve as building blocks to design a feasible, affordable and portable tool to improve the assessment and treatment of neurological disorders. Key challenges are (1) to identify these biomarkers, and (2) to reliably estimate them using automatic systems. A common and straightforward method is to train people to evaluate biomarkers based on auditory perception. However, this method is subjective and prone to ignore subtle acoustic cues. Advances in speech processing provide an alternative solution for this problem. *Automatic Speech Recognition* (ASR) systems are able to automatically recognize the input speech and capture the timing boundaries at different levels, such as words, syllables and phonemes. We can automatically estimate corresponding biomarkers, relying on the detection result provided by the ASR system. Speech biomarkers can contribute with objective and precise metrics without personal bias, providing valuable information for clinicians or speech therapists.

Over the last couple of years, our groups have collaborated to develop stand-alone applications to detect vocal acoustic biomarkers indicative of neurological disorders [2]. These applications are implemented on portable devices (iOS), and are designed to work regardless of Internet accessibility. As a result, all the computation is done on the device, not in the cloud. We have designed a data collection protocol with seven brief tasks, for which we have developed specialized ASR systems. One of them is the *diadochokinetic* (DDK) test consisting of speaking repetitions of "PA-TA-KA", as fast as possible. This popular task has been extensively used in previous work to study the relationship between neurological disorders and DDK rate [3–10]. This task poses important challenges for ASR systems due to the error patterns made by individuals, including mispronunciation, alternating syllable order, skipping syllables, replacing syllables, and restarting the sequences. These errors are also informative of neurological disorder, so it is important to identify them.

Under the constraint that the computation has to be done on the portable device, this paper presents simple but effective speech solutions that work remarkably well for this DDK task. We build syllable-based acoustic models, considering out-of-vocabulary words, and speech fillers. Our language model makes our ASR system robust to syllable sequences that are out of order or incomplete. We obtain accurate syllable boundary detection, which allows us to extract interesting candidate biomarkers, such as DDK rate and temporal evolution of the prosodic structure across multiple repetitions of "PA-TA-KA".

http://dx.doi.org/10.21437/Interspeech.2016-789

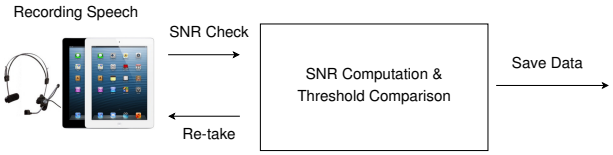

Figure 1: The procedure of data collection

We evaluate this system with data recorded from two important neurological conditions, for which we are collecting extensive recordings. The first condition is *mild traumatic brain injuries* (mTBIs), where we collect data from high school and college athletes participating in sports with high concussion rates. The second condition is *Parkinson's disease* (PD).

2. Related Work

Diadochokinetic (DDK) rate has been used to diagnose and monitor motor speech disorders [11]. This task normally requires subjects to speak repetitions of a single syllable or a group of syllables as fast as possible. By increasing the speech rate, the cognitive and motor function demands increase, especially when the task includes sequential syllables involving different speech articulation changing the configuration of the oral cavity. DDK tasks are able to examine alternate and sequential movements involving specific syllable repetitions, providing valuable tools in neurological disorders diagnosis [12, 13]. High level features such as *sequential motion rate* (SMR), *alternating motion rate* (AMR), temporal variation of *fundamental frequency* (F0), and formant information (e.g. F1, F2) can be extracted from DDK tests and used as standard biomarkers for clinical diagnosis.

The sequence of syllables "PA", "TA", and "KA" are normally used to study the relationship between neurological disorders and DDK rate [3–10]. This sequence requires the subjects to use the lips, the tip of the tongue, and the soft palate. These articulations involve the front, middle, and back parts of the mouth respectively, providing a valuable test to assess motor speech skill. The combination of these three syllables, "PA-TA-KA", in testing oral motor skills and neurological disorders has proven to be more efficient than single syllable tests, because it requires more complex speech articulation, increasing cognitive demands [14–16]. Thus, we decided to investigate simple, but effective solutions to improve the performance of our system utilizing this task. This paper reports the proposed solution for this DDK task.

3. Task Design and Data Collection

This section briefly describes our current protocol and iOS application for data collection. Figure 1 describes the flowchart of our system. We also briefly describe the data collection effort. We describe in detail these aspects in Poellabauer et al. [2].

3.1. Task Design and Application for Portable Devices

The key goal of our effort is to create an application, which is able to derive reliable biomarkers of motor speech disorders using only few minutes of recordings. After multiple iterations in the design of our protocol, we defined seven brief tasks which are listed in Table 1. Task 1 required the subject to speak multisyllabic words, capturing deficits in range, accuracy, and speed of movement during speech production. In Task 2, the subject read the sentence three times placing emphasis on "put", "book" and "here", respectively, capturing restricted ability to produce stress. Task 3 required the subject to accurately read a given sentence. Tasks 4 and 5 were DDK tasks including repeating a single syllable (AMRs). Task 6 corresponded to the "PA-TA-KA" task (SMRs), which is the focus of this paper. Finally, task 7 required the subject to sustain the /ah/ vowel for at least 5 seconds, capturing muscle tone and steadiness of the tone. The entire recordings took between 2 and 3 minutes to complete.

Table 1: Description of speech tasks for the mTBI data collection. A modified version is used for the PD data collection.

ID	Task
1	Participate, Application, Education, Difficulty, Congratulations, Possibility, Mathematical, Opportunity
2	Put the book here
3	We saw several wild animals
4	PA
5	KA
6	PA-TA-KA
7	AAAHHH

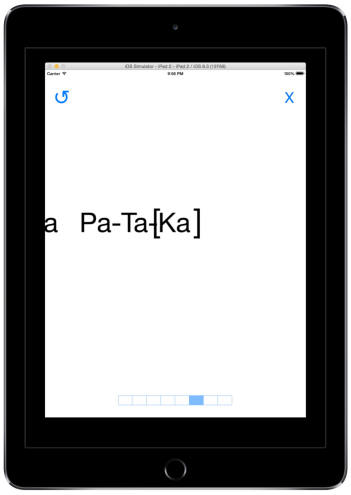

Figure 2: Interface of the iOS application for data collection.

We designed an iOS application for iPads which collected the selected tasks the order shown in Table 1. Figure 2 shows the interface for the task "PA-TA-KA". The application displays the content that the subject is required to read for a fixed duration of 5 seconds. There is a fixed square bracket in the middle of the screen highlighting the text that the user is expected to read. The text moved from one side to the other. A SHURE SM-10 noise cancellation microphone was attached to the iPad, since this external microphone provided higher *signal-to-noise ratio* (SNR) than the iPad's internal microphone [2] (Figure 1). The application is intended to be used in realistic scenarios (e.g., in hospitals, sports arenas, etc.), so we carefully control the noise level. We estimate the SNR in real time and when the noise is above an acceptable level, we ask the subject to move to a quiet room and restart the data collection. The audio was originally recorded at 44.1 kHz, 16 bit, with mono channel. We downsample the audios to 16 kHz for building our "PA-TA-KA" ASR system.

3.2. Data Collection

Our investigation of speech biomarkers is targeted to two neurological disorders: mTBIs and PD. For mTBIs, we collected data from more than 2500 youth athletes participating in sports with high concussion rates from 47 schools in the Midwest (Illi-

Table 2: Summary of subjects in concussion and PD sets.

Set	Total	Female	Male	Age		
				mean	max	min
Concussed	95	16	79	17.5	24	14
Non-Concussed	485	87	398	16.4	22	14
Total	580	103	477	16.6	24	14
PD	7	4	3	65.6	82	57
Non-PD	10	7	3	54.1	76	23
Total	17	11	6	58.5	82	23

nois, Indiana, Michigan) and Pennsylvania. We collected baseline recordings at the beginning of either the academic year or the athletic season, assuming they were in healthy condition. The subjects provided information about age, gender, previous concussion events and other medical records. We repeated the protocol immediately after the athletes finished their competitions. We were only able to collect 580 of these athletes, where 95 of them reported concussion symptoms. Table 2 provides information about this corpus.

Recently, we have started a new project to analyze speech recordings from individuals with PD. This is an ongoing effort where we are collecting speech recordings from PD patients and their spouses (where the spouses serve as control subjects). While the test for PD is slightly different from the test for mTBI, the PA-TA-KA test is unchanged. For this study, we use 10 typical healthy people, and 7 subjects with PD (Table 2).

The recordings for the PA-TA-KA test were manually transcribed, including the timing information for the syllable boundaries. Currently, we are still processing the data, so this paper uses 204 transcribed audios from the concussion set and 17 transcribed audios from the PD set. Each of them have a duration of 5 seconds, including on average 9 repetitions of "PA-TA-KA".

4. Building ASR solutions for PA-TA-KA

Over the last five years, ASR systems built on *deep neural networks* (DNNs) have achieved state-of-the-art performance compared to traditional systems built on *Gaussian mixture models* (GMM) and *hidden Markov models* (HMM) [17, 18]. DNN systems train millions of parameters relying on large databases [19,20]. The computation and memory constraints of these systems is intense, so current ASR solutions for portable devices such as Siri (Apple), Cortana (Microsoft), Alexa (Amazon) perform the computation in the cloud [21]. While there is a growing interest on running all the computation of these ASR systems on portable devices [21, 22], we decided to use a GMM-HMM framework. In particular, we use Pocketsphinx [23], which can be easily incorporated into iOS applications.

We designed simple ASR solutions for all the tasks listed in Table 1 achieving adequate performance, except for the "PA-TA-KA" task, which is the most difficult task from an ASR perspective. Subjects normally make mistakes, including inserting, replacing, or deleting syllables. As participants are asked to utter the syllable sequence as fast as they can, they tend to produce syllables out of order, with poor pronunciations. The participants often laugh and speak out-of-vocabulary words. We address these challenges by building task specific language and acoustic models. Since this study analyzes pre-recorded speech, the processing is performed off-line. However, the ASR system can be easily adapted to work in real time.

4.1. Task-Specific Acoustic Model

We built our ASR system using HMMs, where the acoustic models are trained with GMMs. Instead of building phoneme models, as conventional ASR systems, we created three syllable models for "PA", "TA" and "KA", respectively. Notice that the three syllables have the same vowel and the key difference is on their consonant (bilabial, alveolar, velar). By building syllable models, we capture the consonant-vowel transitions at the HMM level, creating more robust models. These HMMs are built in Pocketsphinx with three states using a left-to-right topology. The GMMs are built with four mixture components.

Our acoustic features correspond to *Mel Frequency Cepstral Coefficients* (MFCCs), which are extracted with Pocketsphinx using a 25ms Hamming window, which is shifted with step size of 10 ms (e.g., 100 frames per second). We use 13 MFCCs where we concatenate the first and second order differences to include temporal information, creating a 39 dimensional feature vector. We normalize the feature vector using mean shift normalization at the utterance level.

We cope with out-of-vocabulary words by creating a background model. We use the CRSS-4ENGLISH-14 corpus [24] for this purpose, where we group all other syllables together. This background model is effective in capturing instances where the participants do not correctly complete the task. We build other models for laughing, clearing throat, coughing, smacking lips, and silence. Our silence model is trained with samples from the concussion set (training partition). We obtain samples for laughing, clearing throat, coughing and smacking lips from the CRSS-4ENGLISH-14 corpus, which has transcriptions for these fillers. We use these samples to build the models for the fillers.

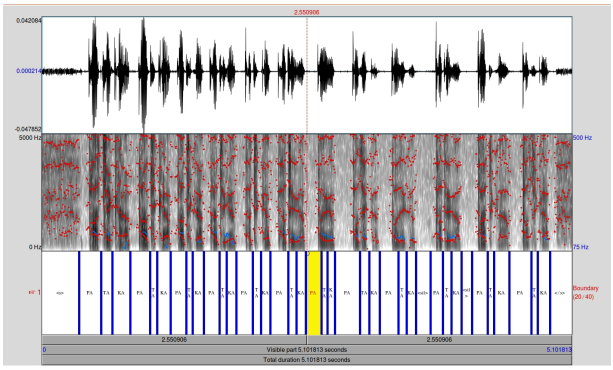

Figure 3: Example of a recording. The Figure shows the speech wave, spectrogram (with formants), and syllable segmentation.

4.2. Language Model

During the DDK task, the subjects have to repeat "PA-TA-KA" as fast as possible. Frequent errors include altering the order of the syllables (e.g., "PA-KA-TA"), missing syllables (e.g., "TA-KA"), or repeating syllables ("PA-PA-TA"). We address these challenges with the language models. Instead of using a strict, fixed grammar, we rely on a trigram language model were we allow the system to account for wrong syllable sequences. We especially consider the syllable errors observed during training, which are used to estimate the transition probabilities. We also consider uni-grams and bigram to account for new sequences of target syllables not observed during training.

5. Experimental Evaluation

Figure 3 shows a speech recording with the results provided by the proposed ASR system. The Figure shows the speech wave, spectrogram (with formants), and syllable segmentation. We

Table 3: The ASR performance in terms of SER and boundary detection. Results are for the concussion ("Con") and PD ("PD") sets.

Set	Conditions	SER[%]	Boundary Detection		
			Pre	Rec	F
Con	Concussed	2.4	0.92	0.48	0.63
	Non-Concussed	3.5	0.91	0.48	0.63
PD	PD	7.9	0.82	0.46	0.59
	Non-PD	6.2	0.85	0.46	0.60

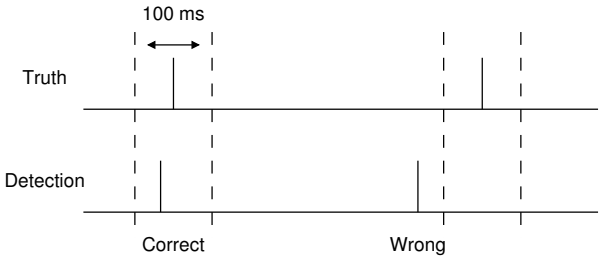

Figure 4: Evaluating boundary detection performance. We use a tolerance window of 100ms around the ground truth label.

evaluate the proposed ASR system training the models with 124 recordings from the concussion set. Each of them have multiple repetitions of PA-TA-KA. All these recordings correspond to individuals without concussion. We test the models with the remaining data: 80 sentences from the concussion set, where 27 presented concussion symptoms; and, 17 sentences from the PD set, where seven were recorded by individuals with PD. We evaluate the accuracy of our system in estimating the syllables and their boundary detection, which are important aspects for the speech biomarkers that we intend to estimate (Section 5.1).

The first metric is the *syllable error rate* (SER) reported in Table 3 for the concussion and PD sets. The system performs well reaching levels of SER under 4% for the concussion set and under 7% for the PD set. For the concussion set, our ASR performs well for both concussed and non-concussed individuals. The performance is around 1.1% better for concussed individuals, since they have less fillers. Due to the low number of recordings from the PD set, we cannot draw conclusive insights. The SER is about 3-4% worse. There is a clear difference in the subject's age, so we are planning to retrain the acoustic models under matched age conditions.

The second metric is the accuracy in detecting the syllable boundary. We set a 100ms window tolerance around the ground truth. For example, if the labelled boundary is located at 2s, a correct detection should lie between 1.95 to 2.05 seconds. Otherwise, it is considered a wrong detection (Figure 4). We estimated the precision ('Pre.'), recall ('Rec.') and F-score ('F'). Table 3 lists the performance, which is consistent across conditions for the concussion and PD sets. We observe precision rates above 0.9 for the concussion set, and 0.8 for the PD set. The detected boundary are mostly correct. However, there are boundaries that we are not detecting, limiting our recall rates.

5.1. Target Speech Biomarkers

The proposed ASR system can be used to extract speech biomarkers that are discriminative of neurological disorders. We can derive various relevant measures, including (1) the number of repetitions of correctly uttered "PA-TA-KA", (2) the average DDK rate, defined as the number of syllables per second, (3) the average DDK period (inversely related to DDK rate), (4) the standard deviation of DDK rate, and (5) the coefficient of variation in DDK period, defined as the degree of variation in the period. With the syllable boundaries, we can also analyze the prosodic structure of consecutive repetitions of "PA-TA-KA". A comprehensive description of these biomarkers is discussed in Poellabauer et al. [2].

Table 4: Estimating candidate speech biomarkers. The table reports the MAD scores for "PA", "TA" and "KA" and complete sequences of correctly uttered "PATAKA". Results are for the concussion ("Con") and PD ("PD") sets.

Set	Conditions	PA	TA	KA	PATAKA
Con.	Concussed	0.20	0.08	0.12	0.32
	Non-Concussed	0.27	0.20	0.24	0.73
PD	PD	0.75	0.50	0.38	0.75
	Non-PD	0.25	0.38	0.25	0.63

As an example, we estimate the number of "PA", "TA", and "KA" syllables uttered per sentence. We define the *mean absolute difference* (MAD) score estimated in Equation 1, where N_{true} is true number of syllables, $N_{detection}$ is the detected number of syllable, and L is the number of recordings in the testing set. For the concussion set, we observed errors below 0.3 which implies that we can reliably estimate the actual number of syllables for each of the target syllables. We observe similar performance for "PA", "TA", and "KA". We included the results for PD set, although the number of recordings is limited. Finally, we estimate the number of "PA-TA-KA" sequences which were correctly spoken. We do not count cases with missing or incorrect parts such as "PA-TA", "PA-KA", "PA". The system can robustly estimate the number of correctly spoken "PA-TA-KA", showing errors less than 0.75 (less than 1 count per recording, which usually includes around 9 repetitions). These results show the feasibility of using the proposed application to derive biomarkers of neurological disorders.

$$MAD = \frac{\sum_{i=1}^{L} |N_{true} - N_{detection}|}{L} \quad (1)$$

6. Conclusions

We presented a task-specific ASR system for the popular DDK test consisting of repetitions of syllables "PA-TA-KA". The proposed approach addresses the key challenges associated with this task, providing a flexible tool for portable devices, where the computation is on the device, not in the cloud. The ASR system robustly detects the target syllables and their boundaries. We demonstrated how this system can be used to estimate speech biomarkers. In particular, we robustly estimated, for a given recording, the number of repetitions for each of the target syllables uttered by the participant, and the total number of correctly uttered sequences of "PA-TA-KA".

We are currently collecting more data from patients diagnosed with PD. We are using the models to automatically derive speech biomarkers that can reveal possible early indications of neurological disorders. We expect to build a stand-alone application for iOS, to record and process PATAKA recordings. From a system perspective, we are looking for post processing approaches to improve the recall rates of the syllable boundary detection. We are exploring alternative methods for syllable boundary segmentations from spectrograms [25]. Finally, we plan to retrain the models for the PD set after collecting enough recordings. We will use data from healthy individuals similar in age, to reduce the existing age mismatch in our current study.

7. References

[1] H. Halpern and R. Goldfarb, *Language and motor speech disorders in adults*. Burlington, MA, USA: Jones & Bartlett Learning, March 2012.

[2] C. Poellabauer, N. Yadav, L. Daudet, S. Schneider, C. Busso, and P. Flynn, "Challenges in concussion detection using vocal acoustic biomarkers," *IEEE Access*, vol. 3, pp. 1143–1160, August 2015.

[3] S. Sheinkopf, J. Iverson, M. Rinaldi, and B. Lester, "Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder," *Autism Research*, vol. 5, no. 5, pp. 331–339, October 2012.

[4] Y. Wang, R. Kent, J. Duffy, and J. Thomas, "Dysarthria associated with traumatic brain injury: Speaking rate and emphatic stress," *Journal of Communication Disorders*, vol. 38, no. 3, pp. 231–260, May-June 2005.

[5] A. Goberman, M. Blomgren, and E. Metzger, "Characteristics of speech disfluency in Parkinson disease," *Journal of Neurolinguistics*, vol. 23, no. 5, pp. 470–478, September 2010.

[6] O. Geman, "Data processing for Parkinson's disease: Tremor, speech and gait signal analysis," in *E-Health and Bioengineering Conference (EHB 2011)*, Iasi, Romania, November 2011, pp. 1–4.

[7] B. Plassman, K. Langa, G. Fishe, S. Heeringa, D. Weir, M. Ofsteda, J. Burke, M. Hurd, G. Potter, W. Rodgers, D. Steffens, R. Willis, and R. Wallace, "Prevalence of dementia in the United States: the aging, demographics, and memory study," *Neuroepidemiology*, vol. 29, no. 1-2, pp. 125–132, November 2007.

[8] L. Goldstein, A. Fisher, C. Tagge, X. Zhang, L. Velisek, J. Sullivan, C. Upreti, J. Kracht, M. Ericsson, M. Wojnarowicz, C. J. Goletiani, G. M. Maglakelidze, J. A. Moncaster, N. Casey, R. D. Moir, O. Minaeva, C. J. Nowinski, R. A. Stern, R. C. Cantu, J. Geiling, J. K. Blusztajn, B. L.Wolozin, T. Ikezu, T. D. Stein, A. E. Budson, N. W. Kowall, D. Chargin, A. Sharon, S. Saman, G. F. Hall, W. C. Moss, R. O. Cleveland, R. E. Tanzi, P. K. Stanton, and A. C. McKee, "Chronic traumatic encephalopathy in blast-exposed military veterans and a blast neurotrauma mouse model," *Science Translational Medicine*, vol. 4, no. 134, pp. 1–16, May 2012.

[9] J. Langlois, W. Rutland-Brown, and M. Wald, "The epidemiology and impact of traumatic brain injury: a brief overview," *Journal of Head Trauma Rehabilitation*, vol. 21, no. 5, pp. 375–378, September/October 2006.

[10] S. Skodda and U. Schlegel, "Speech rate and rhythm in Parkinson's disease," *Movement Disorders*, vol. 23, no. 7, pp. 985–992, May 2008.

[11] S. Fletcher, "Time-by-count measurement of diadochokinetic syllable rate," *Journal of Speech, Language, and Hearing Research*, vol. 15, no. 4, pp. 763–770, December 1972.

[12] P. Williams and J. Stackhouse, "Diadochokinetic skills: Normal and atypical performance in children aged 3-5 years," *International Journal of Language & Communication Disorders*, vol. 33, no. Supp 1, pp. 481–486, 1998.

[13] J. Yaruss and K. Logan, "Evaluating rate, accuracy, and fluency of young children's diadochokinetic productions: a preliminary investigation," *Journal of Fluency Disorders*, vol. 27, no. 1, pp. 65–86, March 2002.

[14] M. Padovani, I. Gielow, and M. Behlau, "Phonarticulatory diadochokinesis in young and elderly individuals," *Arquivos de Neuro-psiquiatria*, vol. 67, no. 1, pp. 58–61, March 2009.

[15] P. Sörös, L. Sokoloff, A. Bose, A. McIntosh, S. Graham, and D. Stuss, "Clustered functional MRI of overt speech production," *NeuroImage*, vol. 32, no. 1, pp. 376–387, August 2006.

[16] S. Ghosh, J. Tourville, and F. Guenther, "A neuroimaging study of premotor lateralization and cerebellar involvement in the production of phonemes and syllables," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 5, pp. 1183–1202, October 2008.

[17] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, January 2012.

[18] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, January 2012.

[19] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[20] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[21] R. Prabhavalkar, O. Alsharif, A. Bruguier, and I. McGraw, "On the compression of recurrent neural networks with an application to LVCSR acoustic modeling for embedded speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5970–5974.

[22] X. Lei, A. Senior, A. Gruenstein, and J. Sorensen, "Accurate and compact large vocabulary speech recognition on mobile devices," in *Interspeech 2013*, Lyon, France, August 2013, pp. 662–665.

[23] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, Toulouse, France, May 2006, pp. 185–188.

[24] M. Nandwana, H. Boril, and J. Hansen, "A new front-end for classification of non-speech sounds: A study on human whistle," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1982–1986.

[25] O. Kalinli, "Syllable segmentation of continuous speech using auditory attention cues," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 425–428.