SYLLABLE LEVEL FEATURES FOR PARKINSON'S DISEASE DETECTION FROM SPEECH

Sevada Hovsepyan and Mathew Magimai.-Doss

Idiap Research Institute, Centre du Parc, Rue Marconi 19 CH - 1920 Martigny Switzerland

ABSTRACT

Early detection of Parkinson's disease (PD), one of the most common neurodegenerative diseases, is crucial for successful treatment and symptom management. In this study, we propose a novel approach inspired by neurocomputational models of speech perception, for PD detection from speech samples. Our proposal emphasises the importance of acoustic/linguistic markers to extract features at the syllable level, in contrast to conventional methods that extract features at the frame or state level. Through the use of syllable-level features (SLF), we successfully identify PD in recorded speech samples. Remarkably, the results not only match but potentially exceed the effectiveness of traditional feature sets used for this purpose. We hope that the proposed approach will provide a new basis for integrating linguistic insights into the identification of speech-related diseases.

***Index Terms*—** syllable-level-features, Parkinson's disease, SVM, language disorder, classification

1. INTRODUCTION

Parkinson's disease (PD) is a prevalent neurodegenerative condition that affects many patients, impacting both motor (e.g. motor planning, tremor, bradykinesia) and non-motor systems (cognitive, gastrointestinal, neuropsychiatric) [1, 2]. These systems also affect how PD patients articulate speech, which is typically described as sad, monotonic, slow, disconnected, etc [3]. Therefore, in recent years, several studies have been conducted on the automated detection of Parkinson's disease based on speech articulation [4, 5, 6].

The primary speech feature of Parkinson's disease (PD) patients is phonation impairment, with articulation being the second most affected subsystem [7]. Therefore, sustained vowel recordings are traditionally used to evaluate phonation impairment, and diadochokinetic (DDK) exercises are conducted to assess articulatrion impairment. When assessing PD's effects on speech over extended timescales, prosody analysis is frequently included. However, to conduct a full analysis of speech in PD patients, it is essential to consider all these factors, as shown in [7]. In their study the authors undertake an acoustic evaluation of the speech of PD patients, specifically exploring their phonation, articulation, and prosody. The speech corpus utilized in this study consists of recordings from 46 individuals who are Czech natives. Among these participants, 23 individuals were identified as having early stage idiopathic PD by neurologist experts, while the remaining 23 individuals were healthy controls (HC). The collection of recordings encompasses the DDK task, sustained phonation of the vowel /i/, aloud readings, and monologues. Traditional features such as formant frequencies, articulation rate, pause characteristics, and voice intensity, coupled with non-standard attributes like relative intensity range variations and spectral distance variations, were adopted to demonstrate that 78% of early untreated PD patients manifest compromised vocal functionality affecting speech subsystems such as phonation, articulation, and prosody, oftentimes in a combination.

In their study, Bayestehtashk et al. 2014 [5] examined the automated assessment of PD severity through speech analysis. The corpus comprised 168 patients at varying stages of the disease who were all native English speakers. The data collected included sustained phonation of the vowel /a/, DDK utterances and reading aloud text. The ComPareE2010 (Computational Paralinguistic Challenge) feature set from INTERSPEECH2010 [8] was used for classification. According to the findings, assessing the severity of the disease is best achieved through reading text and DDK tasks.

In their study, Orozco-Arroyave et al. 2014 [9] investigate various acoustic measures derived from sustained vowel recordings in the PC-GITA database [9]. The analysis comprises various acoustic measures, including the first two formants, pitch, jitter, shimmer, vowel articulation index, triangular vowel space area (tVSA), and three novel tVSA-based metrics. The accuracy of pitch variation measurements is deemed crucial amongst these features. Furthermore, the integration of vocal tract shaping and voice quality attributes resulted in an advancement of the outcomes, attaining an 81.3% accuracy rate of PD classification task.

A recent study [6] utilised a convolutional neural network to investigate the effectiveness of different speech signal segments in automatically detecting PD. The speech corpus included recordings from 268 Lithuanian natives who pronounced phonetically balanced four-word sentences. The feature sets comprised of Mel-frequency spectrograms, first and second derivatives, and other feature maps. Image interpolation was employed to obtain a fixed-length spectrogram. The study demonstrates that dividing speech into segments and fusing the classification decisions from the segments results in better PD detection than classification based on the sentences.

In this study, we present a novel approach for the automatic detection of PD from speech. Our method draws inspiration from neurophysiologically plausible models of speech perception that focus on syllables [10, 11]. When sound enters the ear, it travels through the cochlea and is decomposed into frequency components based on which hair cells in the cochlea were activated. During the process of speech perception, information consistently enters the cortex, and a crucial step for subsequent processing involves segmenting this information into syllabic units [12]. As such, syllables are conventionally considered to be the basic blocks of both speech perception and production [13]. Therefore, we suggest that speech impairment related to PD can be detected at the syllable level. Furthermore, in line with [6], we recommend that continuous speech signals are segmented into syllables and syllable-level features (SLFs) are extracted, which may be more advantageous for PD detection.

The central concept of the proposed methodology is that PD

This work was funded by the Swiss National Science Foundation through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. 40B2 0 194794/1)

affects motor control, leading to impaired speech articulation, including difficulties in pronouncing syllables. Two types of syllable segmentation can be considered. Valley-based segmentation, which entails syllable onset segmentation, concentrates primarily on the information concerning vowel articulations. Thus provides information on phonation impairments. On the other hand, peak-based segmentation centres on consonant information, which requires more refined motor control and can be more challenging for PD patients. Taken together, we hypothesise that peak-based segmentation may yield better PD detection.

The manuscript is divided into the subsequent sections: In Section 2, we present information regarding the neurocomputational models mentioned above. In Section 3, we describe in detail the syllable-level features extraction. In Section 4, we present the dataset and experimental protocols used throughout the study, the baseline features sets and outcomes from conducted experiments. Lastly, in Section 5, we conclude our findings and accentuate the potential prospects of the proposed method.

2. BACKGROUND

In this section, we will provide more details on how some syllable-based neurocomputational models of speech perception represent syllable level information. The basic concept behind syllable level representation is that a syllable can be represented as a pattern in a spectral space, with a series of spectral vectors making up each syllable. One straightforward method to accomplish this is by collapsing the spectral energy of a syllable's time-frequency (TF) decomposition into a fixed number of frequency channels and temporal bins. In several neurocomputational models of speech perception [10, 11, 14], the spectral pattern of a syllable or a single word (taken from models of the auditory periphery [15]) is reduced to 6 channels, by averaging the activity of adjacent frequency bands. The duration of the speech segment is then divided into 8 (arbitrarily chosen) equal parts, and the average activity of each frequency channel within each temporal bin is calculated. Finally, an additional step was undertaken to translate these patterns into the Hopfield space [16]. As a result, each of the 8 6-dimensional spectral vectors of a syllable serve as a global attractor, thus rendering the patterns of different syllables more distinguishable from one another. This representation sufficed to obtain approximately 98.4% accuracy in digit recognition [10] and approximately 40% (with an 8% chance level) accuracy for online syllable recognition [11].

3. PROPOSED METHODOLOGY

In this section, we detail the process for our proposed method, comprising of four steps: 1. the segmentation of continuous speech utterances into syllable-like segments, 2. the calculation of a standardized spectrogram for each syllable, 3. the construction of a feature vector, and finally, 4. the classification of PD from HC based on these feature vectors. Further elaborations of each step are presented below, while Figure 1 illustrates the flowchart of the proposed algorithm.

1. Segmentation: There are several algorithms for automatic syllable onset detection, to name a few [17, 18, 19]. For our current implementation, we have chosen neurophysiologically inspired model utilising theta oscillations and sonority envelope [20]. This approach allows to perform the syllable segmentation either based on syllable onsets (i.e. valleys on the envelope) or syllable nuclei (where typically vowels appear, represented by peaks in the envelope). However, it is worth noting that the proposed approach is agnostic to the segmentation method, and any other syllabification algorithm can, in principle, be used.

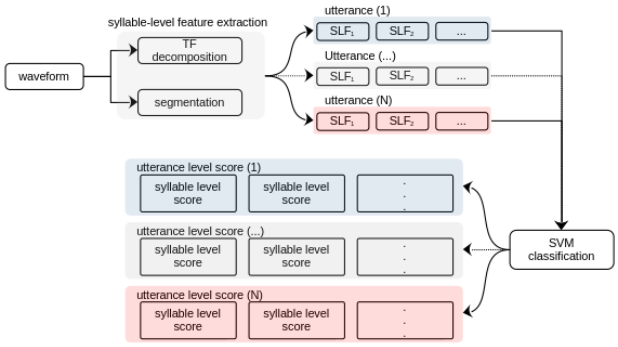

Fig. 1. Flowchart of proposed methodology

2. Standardised spectrogram: The subsequent stage involves calculating a standardised spectrogram - in essence, a time-normalised (fixed temporal bins) and frequency-normalised (same frequency range, number of channels) representation of a syllable. For the frequency domain, we can either follow the procedure described in the Background section (averaging the activity of neighbouring channels and reducing the dimensionality to F - channels), or we can directly compute the time-frequency decomposition with the desired number of channels. Standardisation in the time domain proceeds similarly to the procedure described above: dividing the duration of a syllable's time-frequency decomposition into T equal temporal bins and calculating the average energy of each frequency channel in each bin. As a result, each syllable is presented as a standardised frequency-time matrix (FxT). It should be noted that the number of channels and temporal bins is flexible and can be adjusted to suit a specific task.

3. Syllable level features: Next, for each syllabic segment the feature vector is constructed from the standardised spectrogram by transforming the FxT matrix into a 1x(FxT) vector. Consequently, each speech utterance is presented by a series of feature vectors (as many as detected syllabic segments) that can be utilised for classification. The feature vector of each syllabic segment is used as a data point for the classification procedure.

4. Classification: Finally, the extracted features vectors from each syllabic segment is used for classification to distinguish between PD and HC conditions.

4. RESULTS AND EXPERIMENTAL PROTOCOLS

This section comprises four subsections. In subsection 4.1, we provide information about the dataset used, baseline feature sets, evaluation, and experimental protocols. Subsection 4.2 focuses on the impact of the dimension of the feature vector (specifically the number of frequency channels) on the classification results and its comparison with baseline feature sets. Additionally, other approaches for the time-frequency decomposition are explored. Subsection 4.3 is dedicated to examining the syllable segmentation process and investigating whether peak-based segmentation, which centres on the information surrounding consonant articulation, is more beneficial for the detection of PD compared to valley-based segmentation. It could be argued that consonant articulation is more challenging for patients with PD, as it requires more refined motor control - thus peak based segmentation might be more feasible for PD detection.

We also examine whether the segments must be based solely on syllables. Additionally, in subsection 4.4, we assess the feasibility of applying the suggested methodology to unseen syllables and other speech types (e.g., monologues).

4.1. Dataset and protocols

The dataset used in this study consists of diadochokinetic (DDK) utterances from the PC-GITA speech corpus [9]. The corpus includes recordings of 50 individuals with PD, comprising 25 men with a mean age of 62.2±11.2 years and 25 women with a mean age of 60.1±7.8 years, as well as 50 HC, consisting of 25 men with a mean age of 61.2±11.3 years and 25 women with a mean age of 60.7±7.7 years. All recordings were sampled at 44.1 kHz with 16-bit resolution. Recordings are available for each participant for six different DDK tasks, as listed in the first column of Table 1. Thus, there are 100 recordings for each DDK task, consisting of 50 PD and 50 HC conditions.

We applied the leave-one-subject-out method for each DKK-task, excluding a single utterance as a test case for each fold. The remaining utterances' syllables were used for training the classifier. Our study employed a Support Vector Machine with a cubic kernel to perform classification without hyperparameter optimization, although other classifiers could be used. Each syllable's feature vector was entered as a separate data point for classification. Hence, the AUC (area under the receptive-optimized-curve) along with the corresponding classification performance score is obtainable at the syllable level. However, scores at the utterance or state level can also be obtained by aggregating the scores from related syllable segments (e.g. by arithmetic mean, reported as utterance/state level AUC).

As a baseline features sets we have selected ComParE2016 [21] and eGeMAPSv2 [22]. The ComParE2016 feature set comprises 6373 features, which were also utilized in INTERSPEECH2015 for assessing the severity of PD [23]. Consequently, it serves as a dependable benchmark for validating our proposed approach. The eGeMAPS, or extended Geneva Minimalistic Acoustic Parameter Set, consists of 88 carefully selected features [22]. Its purpose is to provide a standard baseline for speech emotion recognition experiments. We suggest that the inclusion of eGeMAPS as a reliable baseline feature set will enhance the validation of our approach, due to the potential perception of sadness, monotony, or depression in the speech of Parkinson's disease patients. [3]. State-level baseline features were calculated using the OpenSMILE toolbox [24] for each utterance in each DDK task. The AUC values derived from these feature sets to be compared with the AUC values at the utterance level of the proposed method.

4.2. Impact of the number of frequency channels and comparison with the baseline

In this subsection, we examine how the number of channels in a standardised spectrogram, and thus the dimension of the SLF vector, impacts classification results. Additionally, we tested the suitability of more conventional time-frequency decomposition, such as Mel spectrogram or Short-Term Fourier Transform (STFT), in comparison to neurophysiologically plausible auditory spectrogram [15] for the current task.

Following the procedures described in the Section 2, we started with SLF calculated on the valley-based segmentation (Figure 2) and the auditory spectrogram. Table 1 presents the classification results for each DDK task for SLF based on different frequency channel numbers. Our proposed approach outperforms chance-level (50%) performance for all tasks. Also, in agreement with [6], the classification results at the utterance level are superior to those at the syllable level (hereafter the reported results would be based on the utterance level score, if not specified otherwise). On average, the proposed model achieves superior results based on 22 channel SLF representations compared to the average score based on ComParE2016 by approximately 7%, but underperforms the average score of the eGeMAPSv2 feature set by approximately 4%.

task	6 ch.	22 ch.	46 ch.	ComParE	eGeMAPS
papapa	64.83	69.15	67.67		
	72.08	74.12	72.6	68.4	77.66
tatata	62.1	65.16	62.4		
	64.32	65.12	63.76	40.2	73.92
kakaka	73.2	72.74	71.33		
	77.96	75.84	75.64	58.54	70.74
pataka	64.66	67.88	67.07		
	73.68	77.08	74.05	74.82	76.82
petaka	59.07	63.46	62.04		
	67.28	68	66.04	76.56	83.32
pakata	63.23	65.18	63.88		
	70.24	74.08	69.84	73.48	74.72
mean	64.52	67.26	65.73		
	70.93	72.37	70.32	65.33	76.2

Table 1. SVM classification results based on SLF extracted from the auditory spectrogram. For each DDK-task the first row indicates the syllable-level AUC, while the second row represents the utterance/state-level AUC. Column indicates the number of channels used to extract SLF. The last two columns indicate the classification results based on the baseline feature sets.

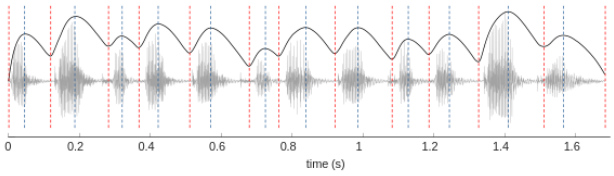

Fig. 2. Segmentation based on valleys (dashed red lines) or peaks (dashed blue lines) of the sonority envelope (black solid line).

Furthermore, in addition to the auditory spectrogram, we examined two alternative methods of computing the time-frequency decomposition of utterances, namely the Mel-spectrogram and Short-Term Fourier Transform (STFT). We adjusted the number of nft points and channels/mels (using the LibROSA library [19]) to produce standardized syllable representations with different number of channels. To ensure that the STFT (or Mel-spectogram) have the highest frequency channel at around 10kHz, to produce F=6 channel STFT(Mel-spectrogram), we first generated F=6+2 to generate 8 channel spectrogram and then excluded the last 2 channels for SLF calculation. Same procedure was performed for the other number of frequency channels. The comparison between outcomes derived from the Mel or STFT spectrograms, and those produced by the auditory spectrogram are presented in Figure 3. SLF based on the 46-STFT representations provides the highest classification performance, outperforming the baseline feature sets (by an average of 14.77% over the ComParE2016 feature set and 3.9% over the eGeMAPSv2 feature set). The results reported in the following sec-

task	valleys	peaks	ΔT_1	ΔT_2	ΔT_3
papapa	78.68	81.96	70.92	67.68	67.24
tatata	75.64	78.84	71.68	73.28	66.72
kakaka	77.96	79.24	65.76	58.92	64.68
pataka	83.64	89.04	77.84	83.82	80.56
petaka	80.48	84.88	73.23	70.12	74.84
pakata	84.2	85.44	71.68	75.16	73.88
mean	80.1	83.23	71.85	71.48	71.32

Table 2. SVM classification results based on SLF extracted from 46-channel STFT spectrogram, where ΔT_1 = 25-50ms, ΔT_2 = 50-450ms, ΔT_3 = 300-600ms

tions are therefore based on the 46-channel STFT representations, unless otherwise specified.

Overall, it can be asserted that the proposed approach, which is based on syllable-level features, has merits and provides impressive classification results compared to the baseline feature sets. Moreover, it is worth emphasising that the score at the level of the utterance, which incorporates the classification results of every syllable, generates superior performance.

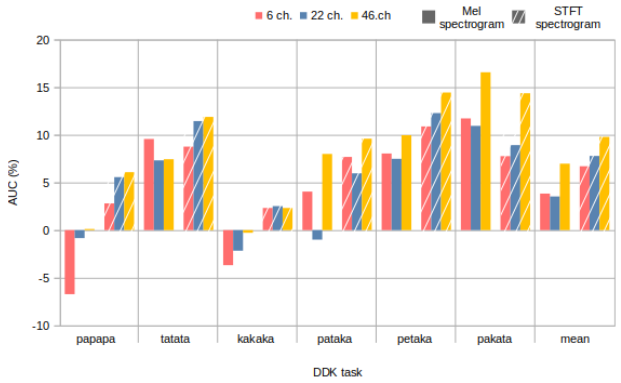

Fig. 3. Performance gain with Mel (solid bars) or STFT (dashed bars) spectrogram versus auditory spectrogram.

4.3. Segmentation: peaks or valleys

All the previous reported results were reported when the segmentation was performed on the local minima (valley) of the sonority envelope, constituting the "syllable" onset. In this section, we investigate 1) whether segmentation methods based on energy peaks (syllable nuclei) would be more advantageous and 2) whether segments should necessarily be based on linguistic/acoustic cues (syllable onset or nuclei). Therefore, we repeated the above-mentioned experiment with peak-based segmentation and random segments (different duration ranges were tested) not necessarily aligned with linguistic/acoustic markers. As can be observed (Table 2), the results are further enhanced with peak-based segmentation, and both syllable-based segmentations outperform segments of similar duration range that lack acoustic/linguistic markers. To summarise, the proposed model's advantage stems from the segments being based on syllabic markers (thus utilizing acoustic/linguistic information), more so if the segment boundaries are based on syllabic nuclei/energy peaks.

tasks	papapa	tatata	kakaka
papapa	81.76	82.84	84.55
tatata	89.76	75.51	84.4
kakaka	87.4	86.75	79.16

Table 3. Cross-task SVM classification results based on 46-channel STFT spectrogram with peak-based syllable segmentation.

duration (s)	3.5	10	12
AUC	65.52	75.68	78.16

Table 4. SVM classification results based on 46-channel STFT spectrogram for monologue extracts. The peak-based segmentation is used.

4.4. Generalization to unseen syllables

All experiments described above were performed on specific DKK tasks. However, in order to test how generalisable our approach is we conducted a cross-task experiment: features extracted from one task (e.g. *papapa*) were used for classification for the other task (e.g. *kakaka*). To ensure that the experiment is speaker-independent, we adapted the LOSO protocol for the cross-task design: the utterance of the "left-out" speaker was excluded from the training set. Moreover, to ensure that the same syllable is not included in both test and train sets, we only used the DDK tasks with single syllable. Results, presented in Table 3, indicate that SLF captures differences between conditions not specific to the DDK task, hence the proposed approach is generalisable to unseen syllables.

Encouraged by these findings, we conducted further tests to ascertain the feasibility of the proposed approach by using monologue extracts from the PC-GITA database. Table 4 presents the results for different extract durations (10 s being approximately the average duration of a DKK task) and suggests that SLF can be applied to broader speech samples.

5. CONCLUSIONS

Overall, the findings suggest that the proposed approach of using a syllable-level feature set for detecting Parkinson's disease from speech samples holds promise. Our results are comparable, if not superior, to classification results based on more typical feature sets like ComParE2016 and eGeMAPSv2 [24, 21, 22]. We have shown that spectrotemporal patterns of syllable-segments, based on STFT, provide sufficient information for distinguishing PD from HC.

Additionally, the results suggest that peak-based segmentation is more effective and supports our initial hypothesis that patients with Parkinson's disease may experience difficulty in articulating consonants. Furthermore, we have noted that the proposed methodology is adaptable and can extract speech-related features applicable to different types of speech, be it a monologue or otherwise.

The proposed approach is versatile in its processing steps and is applicable for detecting speech impairments such as stuttering, dysarthria, and drowsiness in addition to Parkinson's disease. Results indicate that features driven linguistic markers can efficiently detect Parkinson's disease through speech. In the future, further iterations of this study should be undertaken to customize the method for other speech- related disorders.

6. REFERENCES

[1] Jinee Goyal, Padmavati Khandnor, and Trilok Chand Aseri, "Classification, Prediction, and Monitoring of Parkinson's disease using Computer Assisted Technologies: A Comparative Analysis," *Engineering Applications of Artificial Intelligence*, vol. 96, pp. 103955, Nov. 2020.

[2] M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group," *Neurology*, vol. 54, no. 11 Suppl 5, pp. S21–23, 2000.

[3] Konstantinos Sechidis, Riccardo Fusaroli, Juan Rafael Orozco-Arroyave, Detlef Wolf, and Yan-Ping Zhang, "A machine learning perspective on the emotional content of Parkinsonian speech," *Artificial Intelligence in Medicine*, vol. 115, pp. 102061, May 2021.

[4] Juan Rafael Orozco-Arroyave, Elkyn Alexander Belalcázar-Bolaños, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, Tino Haderlein, and Elmar Nöth, "Phonation and Articulation Analysis of Spanish Vowels for Automatic Detection of Parkinson's Disease," in *Text, Speech and Dialogue*, Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, Eds., Cham, 2014, Lecture Notes in Computer Science, pp. 374–381, Springer International Publishing.

[5] Alireza Bayestehtashk, Meysam Asgari, Izhak Shafran, and James McNames, "Fully Automated Assessment of the Severity of Parkinson's Disease from Speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 172–185, Jan. 2015.

[6] Evaldas Vaiciukynas, Adas Gelzinis, Antanas Verikas, and Marija Bacauskiene, "Parkinson's Disease Detection from Speech Using Convolutional Neural Networks," pp. 206–215. Jan. 2018.

[7] Jan Rusz, Jan Rusz, Roman Cmejla, Hana Ruzickova, Evžen Růžička, and Evzen Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, Feb. 2011.

[8] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Interspeech 2010*. Sept. 2010, pp. 2794–2797, ISCA.

[9] Juan Rafael Orozco, Julian D. Arias-Londoño, J. Vargas-Bonilla, María González-Rátiva, and Elmar Noeth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," May 2014.

[10] Izzet B. Yildiz, Katharina von Kriegstein, and Stefan J. Kiebel, "From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems," *PLoS Computational Biology*, vol. 9, no. 9, pp. e1003219–e1003219, Sept. 2013.

[11] Sevada Hovsepyan, Itsaso Olasagasti, and Anne-Lise Giraud, "Combining predictive coding and neural oscillations enables online syllable recognition in natural speech," *Nature Communications*, vol. 11, no. 1, pp. 3117, June 2020.

[12] Morten H. Christiansen and Nick Chater, "The Now-or-Never bottleneck: A fundamental constraint on language," *Behavioral and Brain Sciences*, vol. 39, pp. e62, 2016.

[13] Oded Ghitza, "The theta-syllable: A unit of speech information defined by cortical function," *Frontiers in Psychology*, vol. 4, no. MAR, 2013.

[14] Yaqing Su, Lucy J. MacGregor, Itsaso Olasagasti, and Anne-Lise Giraud, "A deep hierarchy of predictions enables online meaning extraction in a computational model of human speech comprehension," *PLOS Biology*, vol. 21, no. 3, pp. e3002046, Mar. 2023.

[15] Taishih Chi, Powen Ru, and Shihab A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[16] John J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Feynman and Computation*, vol. 79, no. 8, pp. 7–19, Apr. 2018.

[17] Paul Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, Oct. 1975.

[18] R. Villing, T. Ward, and J. Timoney, "Performance limits for envelope based automatic syllable segmentation," in *IET Irish Signals and Systems Conference (ISSC 2006)*. 2006, vol. 2006, pp. 521–526, IEE.

[19] Alexandre Hyafil and Milos Cernak, "Neuromorphic based oscillatory device for incremental syllable boundary detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2015, vol. 2015-Janua, pp. 1191–1195, ISCA.

[20] Okko Räsänen, Gabriel Doyle, and Michael C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, Feb. 2018.

[21] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Interspeech 2016*. Sept. 2016, pp. 2001–2005, ISCA.

[22] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[23] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, J. R. Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Interspeech 2015*. Sept. 2015, pp. 478–482, ISCA.

[24] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, Firenze Italy, Oct. 2010, pp. 1459–1462, ACM.