An Optimized Signal Processing Pipeline for Syllable Detection and Speech Rate Estimation

Kamini Sabu, Syomantak Chaudhuri, Preeti Rao, Mahesh Patil

Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
{kaminisabu,syomantak,prao,mbpatil}@ee.iitb.ac.in

Abstract—**Syllable detection is an important speech analysis task with applications in speech rate estimation, word segmentation, and automatic prosody detection. Based on the well-understood acoustic correlates of speech articulation, it has been realized by local peak picking on a frequency-weighted energy contour that represents vowel sonority. While several of the analysis parameters are set based on known speech signal properties, the selection of the frequency-weighting coefficients and peak-picking threshold typically involves heuristics, raising the possibility of data-based optimisation. In this work, we consider the optimization of the parameters based on the direct minimization of naturally arising task-specific objective functions. The resulting non-convex cost function is minimized using a population-based search algorithm to achieve a performance that exceeds previously published performance results on the same corpus using a relatively low amount of labeled data. Further, the optimisation of system parameters on a different corpus is shown to result in an explainable change in the optimal values.**

Index Terms—**syllable detection, speech rate estimation, signal processing optimization, particle swarm optimization**

I. INTRODUCTION

The syllable-level segmentation of speech is a critical component of speech processing systems with several applications including spoken language assessment for attributes such as speech rate and prosodic fluency. It has been addressed in the research literature over many decades, right up to the current day [1], [2], [3], [4], [5]. While a trained automatic speech recognition system can provide syllable- and word-level segmentations, it has been more common to employ far simpler speech signal processing pipelines to the same end without explicit phonetic segmentation. Further, given that most languages are characterized by syllabic structures comprising vowel nuclei, a method that computes the time variation of a vowel property such as sonority easily generalizes across languages. A vowel is articulated with a relatively open vocal tract configuration making it more energetic than consonants, especially in the vowel formant regions. Therefore performing peak and/or valley detections on a suitably computed band-weighted energy feature versus time can yield the number

and locations of the syllables in an utterance. This is indeed the basis of the popular Mermelstein algorithm for syllable detection, widely deployed to this day [6], [7]. Speech rate is estimated by the number of detected syllables per second averaged over the desired utterance duration.

The performance of the automatic systems is evaluated by comparison with human annotated speech corpora. An algorithmic detection is considered correct if it falls within a tolerance window (in time) of a manually marked syllable nucleus. The following measures then serve to capture the system performance on a test dataset: achieved recall and precision of detections (further combined into F-score) [4], [8], and the accuracy in speech rate estimation via the measured error in the total number of syllables detected over the utterance duration or the correlation between predicted and ground-truth speech rates [4], [3], [9].

A number of parameters are involved in the signal processing pipeline from the acoustic waveform of the utterance to the detected syllabic nuclei. The sub-modules involved are the short-time feature computation, temporal smoothing and, finally, picking the prominent peaks in the smoothened time series. The pipeline involves a set of parameters for each sub-module. A common feature computation is the weighted combination of band-wise short-time energies [3], [10]. Reasonable choices for the processing parameters can be made in the case of short-time spectrum computation (i.e. window and hop sizes are based on speech signal stationarity), as also in the cut-off frequency of the smoothing filter applied to the sonority feature time series (based on the maximum expected syllable rate or fastest speaking speed). The sub-bands are logarithmically spaced according to the auditory model while the number of bands determines the dimensionality of the weight vector. On the other hand, the band-weighting coefficients as well as peak-picking threshold are typically selected heuristically. Such a methodology results in a completely unsupervised approach to syllable detection which, interestingly, has been reported to be competitive with recent deep learning approaches such as a single network of stacked BLSTMs trained on large multilingual corpus [11], [12], [5].

The speech signal processing pipeline mentioned earlier is essentially a semantic abstraction based modular system, unlike neural network based learning systems where the time domain waveform or short-term spectra are input and speech-rate is output. The latter are expected, with their many more degrees of freedom, to require exponentially more data to train and may not generalize well to unseen conditions [13]. We therefore explore here the approach followed by the majority of researchers and practitioners, in the context of the speech syllable detection task, of utilizing prior knowledge to build a pipeline of semantically meaningful sub-modules comprising various signal processing operations while considering further the joint optimization of a small set of critical parameters for given data characteristics.

Given that the choice of sub-module parameters has a significant impact on performance, there have been a few previous attempts at optimizing the parameters over a labeled dataset. Wang and Narayanan [10] exhaustively search the parameter space to scan all possible local maxima of a performance measure in the given range where range and step-size are initialized by Monte-Carlo simulation and a parameter sensitivity analysis. Some optimization approaches avoid the need to explicitly determine the peak-picking threshold by resorting to statistical learning methods that directly learn the speech rate using GMM or HMM modeling [2]. Jiao et al. [3] avoid the peak-picking threshold to set up a convex cost function for speech rate estimation involving integrating a special temporal density function to optimize the sub-bands' weighting coefficients for different dimensionalities of the coefficient vector. Yarra et al. [4] take the approach of modeling the peak shape with a number of parameters and using peak classification (valid versus invalid) based on a feature vector of the estimated peak model parameter values. In a similar manner, Shankar and Venkataraman [14] apply a loose threshold for syllable peak detection but focus on discriminating vowel peaks from consonant peaks using a trained classifier with several features computed on the peak.

In this paper, our aim is to carry out the joint optimization of the system parameters based on a task-specific cost function. We use an optimization technique that is applicable to convex as well as non-convex optimization problems. While learning algorithms optimize the (possibly intractable) task-specific performance measure indirectly via a surrogate loss function, optimization algorithms do a direct optimization[15]. We choose particle swarm optimization (PSO), a non-gradient based algorithm, as it removes the constraint of convexity on the cost function. This is important since the natural objective of our learning task is a non-convex function (such as, for example, the F-score for syllable detection). Other attractive features of PSO are its efficient implementation and its ability to find global optima in multidimensional spaces[16].

In the next section, we describe the system implementation and discuss the optimization of parameters. We present experimental results of the optimized system performance on the TIMIT test corpus [17] in order to compare it with available previously published results of the similar systems

with different approaches to parameter optimization. We also present and discuss the accuracies obtained on a children's oral reading corpus of interest to us in the context of reading fluency assessment [18].

II. SIGNAL PROCESSING PIPELINE

A. Sub-modules and fixed parameters

As mentioned in Section I, the processing pipeline is essentially comprised of the computation of a temporal contour of weighted log sub-band energies followed by low-pass filter smoothing and subsequent peak picking. This is related to the band-wise correlation proposed in [1] and adopted by several others including [3] and [10]. Figure 1 shows a block diagram of the processing chain. Based on the observation by Jiao et al. [3] that 7 sub-bands, derived from the mel scale, provide performance similar to that obtained with 19 sub-bands assuming that the sub-band weighting coefficients are optimized in the same way, we use 7 logarithmically spaced sub-bands in our work.

Figure 1 shows the different stages in the system to detect syllable nucleus locations from the given utterance waveform, normalized to [-1,1] range. All processing is carried out at 10 ms frame interval. A short-time energy based threshold is used to identify the speech frames. The short time window is 20 ms Hamming for the spectrum computation and subsequent sub-band energy computation.
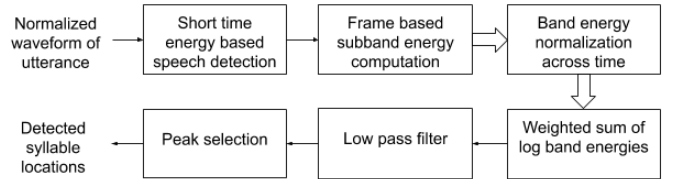

Fig. 1. Signal processing pipeline for syllable detection

In our work, we have considered the 7 frequency bands as provided in Table I. The subbands are defined by a

TABLE I
SYSTEM PARAMETERS WITH SEARCH RANGES FOR THOSE THAT ARE JOINTLY OPTIMIZED BY THE PSO WITH THE GIVEN SETTINGS

Parameter	Value
Window size	20 ms
Frame size	10 ms
Speech energy threshold	-30 dB
Smoothing filter cutoff	7 Hz
Sub-band edges (Hz) (7)	60-370, 370-800,800-1400, 1400-2250,2250-3450, 3450-5130, 5130-7500
Weights search range	-2.0 to 5.0
Peak prominence range	0.01 to 10.0
PSO settings	
No. of particles	50
Maximum no. of iterations	200
ϕ_p, ϕ_g, ω	1.4962, 1.4962, 0.7298

trapezoidal frequency window with 50 Hz transition regions overlapping between bands. We thus obtain 7 band energy values corresponding to each 10 ms frame. The band energy

contours so obtained are individually normalized across the utterance to a maximum value of 1.0 to account for loudness variability. Next, 7 band energy features are computed per frame by taking logarithm of band energies, after limiting very low values to a fixed small positive value. The log energy values are combined by a weighted sum in every frame to obtain a temporal envelope sampled at 10 ms as in equation 1.

$$E[n] = \sum_{i=1}^{7} w_i e_i[n] \qquad (1)$$

where $e_i[n]$ is the log energy feature in frame n of band i and w_i is the corresponding weight.

The temporal envelope so obtained is then smoothened with a low pass filter of cut-off frequency 7 Hz. The cut-off frequency is fixed from the knowledge of speech syllable rates as ranging up to around 7 syllables per sec [19]. The smoothened temporal envelope is searched for peaks (i.e. local maxima) that stand out in a local context. A widely accepted rule is to apply a threshold to each local maximum's "prominence" value, defined as the distance from the nearest adjoining valley [8], [10]. The prominence threshold influences syllable detection performance directly. It is expected to depend on the data spectral characteristics, choice of sub-bands and the sub-band energy weighting coefficients. In this work, we investigate the joint optimization of the weighting coefficients, given a peak prominence threshold, based on a selected task-specific cost function as discussed next.

B. Cost functions for parameter optimization

Performance measures in the context of syllable detection include comparisons between the number of human and automatically detected syllables at the utterance-level, and the precision and recall of automatic detection with reference to human detections. We note that a "correct detection" is defined as that falling within an annotated vowel segment which itself is padded symmetrically to a minimum length of 50 ms if necessary [11]. An automatic detection that does not satisfy this criterion is considered a false alarm. If the end application requires speech rate estimation, on the other hand, the total number of detected syllables per utterance must be compared with the corresponding number of labeled syllables. The correlation coefficient between automatic and ground-truth speech rates (number of syllables/sec) is another useful performance measure provided the utterance duration is long enough for its reliable estimation [9].

Given the flexibility available with PSO in specifying the cost function computationally, we can exploit costs directly related to a performance measure that is eventually important to us in an application without worrying about differentiability. We therefore consider the two performance goals of syllable detection and speech rate estimation separately to obtain the two distinct cost functions below.

1) F-score inverse (based on Precision and Recall in syllable detection)

$$cost = \frac{P + R}{2\,P\,R}, \qquad (2)$$

where

$$P = \frac{\#\text{correctly detected syl}}{\#\text{predicted syl}}$$

$$R = \frac{\#\text{correctly detected syl}}{\#\text{actual syl}}$$

2) Mean absolute error (MAE) for number of syllables per utterance

$$cost = \frac{\sum_{i=1}^{N} |\#syl_{predicted} - \#syl_{actual}|}{N}, \qquad (3)$$

where N = number of utterances.

While the F-score is sensitive to the accuracy in syllable nucleus localization, the MAE relates to variation only in the total number of syllables detected over the utterance duration.

III. EXPERIMENTAL VALIDATION

Comparing performances across different approaches in the literature is difficult in general due to differences in corpora and performance measures. We made our best attempt to identify comparable test scenarios and found published results on TIMIT test set with optimization, if any, carried out on the TIMIT train data set [3], [4] . We also provide results on a children's oral reading corpus created by us for the development of an automatic fluency assessment system. The corpus is named CS dataset (for 'Campus School', where it was recorded).

A. Datasets

TIMIT dataset has 6300 sentences recorded in studio environment. Data is collected from 8 English dialect regions in U.S. and is phonetically transcribed [17]. The utterances have 4 to 20 syllables with duration ranging from 1 to 8 sec. The train and test sets are specified without speaker overlaps. The CS dataset comprises recordings from 10-14 year old Indian children reading aloud the text of short stories in English in quiet conditions. They have different levels of reading skills. The recordings are transcribed at word level. The phone level alignment is obtained next through forced alignment with a state of the art automatic speech recognition system [20]. The story recordings are chunked into sentence groups for the assessment of reading fluency including speech rate. The resulting utterance chunks range in duration between 0.5 s to 15 s. The train-test data distribution is given in Table II. There is no speaker overlap between the train and test sets.

TABLE II
DESCRIPTION OF THE DATA SETS

Dataset	CS data		TIMIT data	
	Train	Test	Train	Test
No. of speakers	16	7	462	168
No. of utterances	2616	1202	4620	1680

B. Parameter Optimization with PSO

Particle Swarm Optimization (PSO) originally proposed by Kennedy and Eberhart in 1995 [21] is widely considered a metaheuristic algorithm well suited to constrained non-convex optimization problems in high dimension space. It operates with a population of candidate solutions ("particles") moving in the specified parameter space, analogous to a swarm of birds. It has a low number of adjustable parameters. It does not require the cost function to be differentiable, or even to be specifiable by an equation. It is essentially applicable to *any* cost function. We use the processing pipeline of Figure 1 with the desired performance measure to compute the cost or fitness function to be minimized in each iteration across the training data set in order to obtain the optimum 7-dimensional weight vector. We see, from equation 1, that the weights are continuous-valued scalars and that the optimal prominence threshold scales linearly with any constant scaling of the weights. Table I lists the PSO parameter settings used in our work based on the common recommended values across diverse PSO applications [22]. Constraints are placed on the parameters via the search ranges specified heuristically in Table I.

C. Evaluation of system performance

A number of standard metrics are available to evaluate the performance of an automatic system on the tasks of syllable detection and speech-rate estimation on a given test dataset as listed below.

- Correlation coefficient between actual number of syllables and predicted number of syllables at utterance level [3], [5].
- SR error rate = Mean of the relative error in number of predicted syllables per utterance [3].
- F-score for syllable detection [4], [7], [23].

We do not report speech rate correlations, expected to be unreliable due to the domination of short spurts (i.e. utterances of duration < 5 s) in the TIMIT dataset [9].

IV. RESULTS AND DISCUSSION

We present results based on the standard performance metrics provided in the previous section on each of the TIMIT and CS test data sets. The training data sets are used in the joint parameter optimization process. In particular, we study the variation in performance with different (i) training data size, (ii) cost functions. Wherever possible, we provide a comparison of our system's performance with available published results on the TIMIT test set. We also investigate the possible corpus dependence of the optimal parameters by comparing the performance achieved on CS test data with parameters that are trained separately on each of the TIMIT and CS training data sets.

Figure 2 shows the variation of relative error in the number of syllables per utterance on the TIMIT test set with an increasing number of TIMIT train sentences used in the optimization, up to the maximum of 4620 sentences. We observe that the error decreases rapidly towards the 100 sentences mark and

then fluctuates in a narrow band. This indicates that a relatively small data set is sufficient to achieve the joint parameter optimization by PSO. This is encouraging in view of the fact that phonetic (or syllable) labeled data sets are expensive to build, and therefore makes the optimization method more usable in practice than might be expected with a deep learning based network. The computation time for PSO optimization on the entire TIMIT training set (4620 sentences) with 100 iterations and 50 particles was limited to within 15 minutes on an i7 CPU with 16 GB RAM. The convergence at relative small training data set size was also observed by Jiao et al. [3] in the context of a convex optimization of weight coefficients for the same task.
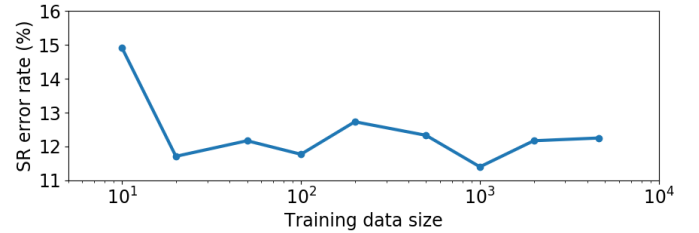

Fig. 2. Performance on TIMIT test set using the 1/F-score cost function versus number of training sentences

Table III shows the results across performance metrics of the proposed system optimized with the two different cost functions as provided in Section II-B. The F-score based cost function targets accurate detection (and localization) of individual syllables. The MAE, on the other hand, is the average error between the actual and predicted number of syllables per utterance, and is not so sensitive to the distinction between correct detections and false alarms. Since speech-rate depends only on the number of detected syllables, we expect MAE to be better suited to optimization for speech-rate estimation. Table III also shows results from previous systems on the same task and same training-testing corpus. For a fair comparison with Yarra et al. [4], we also provide results after a further post-processing step of rejecting peaks with low detected voicing within the parantheses in Table III.

TABLE III
PERFORMANCE ON TIMIT TEST DATA WITH MODEL TRAINED ON 1000 TRAINING INSTANCES OF TIMIT. PERFORMANCE WITH VOICING APPLIED SHOWN IN PARENTHESES.

Method	Corr Coef of #syl	SR error rate (%)	F-score
Jiao et al. [3]	0.842	13.7	-
Yarra et al. [4]	-	-	84.16
Proposed (1/F-score)	0.906 (0.917)	11.4 (11.89)	87.73 (88.29)
Proposed (MAE)	0.890 (0.913)	10.8 (9.94)	86.15 (87.88)

We observe from Table III that the F-score achieved by the proposed system using either cost function is superior to the previously published result based on peak classification [4]. We further note that the proposed joint optimization exceeds the performance of Jiao et al. [3] (errors going down, correlations going up) on all the performance metrics with both

cost functions. As expected, the F-score based cost function also provides better F-score performance while the MAE cost function is better for speech rate estimation. This validates our hypothesis about the benefit of employing a task-relevant cost function for the optimization of system parameters. Further, the analysis of observed syllable detection errors shows that these typically come from short duration vowels (as syllabic nuclei) and syllables devoid of vowels (so that there is no clear valley in the temporal envelope between syllables).

TABLE IV
PERFORMANCE ON CS TEST DATA WITH PARAMETERS OPTIMIZED ON 1000 TIMIT SENTENCES, OR ON 1000 CS TRAIN UTTERANCES, BOTH WITH (1/F-SCORE) AS COST FUNCTION.

Training corpus	Corr Coef of #syl	SR error rate (%)	F-score
TIMIT	0.959	9.9	87.56
CS	0.980	7.3	93.74

Table IV presents results on the CS test data set. With the TIMIT train set used for optimization, we observe better performance on all metrics relative to those obtained on the TIMIT test set as reported in Table III. This may be explained by the generally slower speech rate and clearer articulation by children reading aloud. But, more importantly, we observe the superiority of parameters optimized on training data drawn from the same corpus. This demonstrates the corpus dependence of the optimal parameters. This aspect has been noted by Jiao et al. [3] as well. While the lack of generalization may appear to be a drawback, we keep in mind our previous observation that the amount of data required for optimization to converge is relatively low for any given corpus, making data-based optimization practical in many scenarios. A better understanding of parameter dependence on corpus comes from comparing the optimal sub-band weights determined across the two corpora in Figure 3 . We see that the shapes of the contours are similar with the CS corpus weights being more skewed to the right. This frequency scaling may be explained by the fact
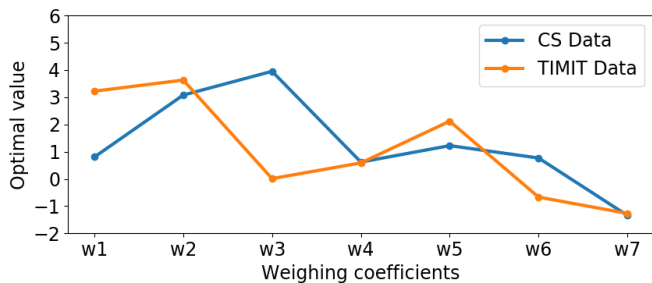

Fig. 3. Optimal weight values obtained by training with 1/F-score cost function on 1000 training utterances of TIMIT data and on CS data. Peak prominence threshold is 2.416 for TIMIT data and 7.341 for CS data.

that the CS data is characterized by higher formant frequencies due to the shorter vocal tract lengths of children. We also note that the weights are smaller in both corpora for the higher frequency bands with negative values in the highest band. This suggests that the features which make syllable nuclei

prominent are, both, the strong presence of vowel formant band energies and the simultaneous absence of high-frequency energy concentration that characterizes consonants such as fricatives and stops.

V. CONCLUSION

We have considered the task of syllable detection in speech signals using different measures of performance. We have demonstrated the data-based optimization of a popular signal processing pipeline using an efficient method, in terms of both computation time and labeled training data size, that places no constraints on the cost function form. Consistent improvements are observed on performance measures of interest based on the appropriate choice of cost function. Future work targets extensions to different language corpora as also to further improvements in performance by incorporating new post-processing methods on the detected peaks.

REFERENCES

[1] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proceedings of ICASSP*, Seattle, USA, 1998.
[2] R. Faltlhauser, T. Pfau, and G. Ruske, "On-line speaking rate estimation using gaussian mixture models," in *Proceedings of ICASSP*, Istanbul, Turkey, 2000.
[3] Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1421–1430, 2015.
[4] C. Yarra, O. Deshmukh, and P. Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," *Speech Communication*, vol. 78, pp. 62–71, 2016.
[5] O. Rasanen, S. Seshadri, J. Karadayi, E. Riebling, J. Bunce, A. Cristia, F. Metze, M. Casillas, C. Rosemberg, E. Bergelson, and M. Soderstrom, "Automatic word count estimation from daylong child-centered recordings in various language environments using language independent syllabification of speech," *Speech Communication*, vol. 113, pp. 63–80, 2019.
[6] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of Acoustic Society of America*, vol. 58, no. 4, pp. 880–883, 1975.
[7] N. Obin, F. Lamare, and A. Roebel, "Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllable," in *Proceedings of ICASSP*, Vancouver, Canada, 2013.
[8] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection," in *Proceedings of ICASSP*, Taipei, Taiwan, 2009.
[9] N. Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.
[10] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.
[11] C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg, and B. Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," in *Proceedings of ICASSP*, Prague, Czech Republic, 2011.
[12] O. Rasanen, S. Shreyas, and M. Casillas, "Comparison of syllabification algorithms and training strategies for robust word count estimation across different languages and recording conditions," in *Proceedings of INTERSPEECH*, 2018.
[13] S. Shalev-Shwartz and A. Shashua, "On the sample complexity of end-to-end training vs. semantic abstraction training," *arXiv preprint arXiv:1604.06915*, 2016.
[14] R. Shankar and A. Venkataraman, "Weakly Supervised Syllable Segmentation by Vowel-Consonant Peak Classification," in *Proceedings of Interspeech 2019*, 2019, pp. Grasz, Austria.
[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[16] S. Sengupta, S. Basak, and R. P. II, "Particle swarm optimization: A survey of historical and recent developments with hybridization perspectives," *Machine Learning Knowledge Extraction*, vol. 1, pp. 157–191, 2018.

[17] J. S. Garofolo and L. D. C. et al., "Timit: Acoustic-phonetic continuous speech corpus," in *Proceedings of the Linguistic Data Consortium*, Philadelphia, PA, USA, 1993.

[18] K. Sabu, K. Nathani, and P. Rao, "Automatic detection of expressiveness in oral reading," in *Proceedings of INTERSPEECH (Show and Tell)*, Hyderabad, India, 2018.

[19] H. Tivadar, "Speech rate in phonetic-phonological analysis of public speech (using the example of political and media speech)," *Journal of Linguistics*, vol. 68, no. 1, pp. 37–56, 2017.

[20] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao, "Automatic assessment of children's L2 reading for accuracy and fluency," in *Proceedings of SLaTE*, Stockholm, Sweden, 2017.

[21] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the International Conference on Neural Networks*, Perth, Australia, 1995.

[22] R. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization," in *Proceedings of the congress on evolutionary computation*, 2000.

[23] B. Ludusan and E. Dupoux, "Automatic syllable segmentation using broad phonetic class information," *Procedia Computer Science*, vol. 81, pp. 101–106, 2016.