Automated Evaluation of Non-Native English Pronunciation Quality: Combining Knowledge- and Data-Driven Features at Multiple Time Scales

Matthew P. Black[1,2,3], *Daniel Bone*[1], *Zisis I. Skordilis*[1], *Rahul Gupta*[1], *Wei Xia*[1], *Pavlos Papadopoulos*[1], *Sandeep Nallan Chakravarthula*[1], *Bo Xiao*[1], *Maarten Van Segbroeck*[1], *Jangwon Kim*[1], *Panayiotis G. Georgiou*[1], *and Shrikanth S. Narayanan*[1,2,3]

[1]Signal Analysis & Interpretation Laboratory, Univ. of Southern California, Los Angeles, CA, USA
[2]Information Sciences Institute, Univ. of Southern California, Marina del Rey, CA, USA
[3]Behavioral Informatix, LLC, Los Angeles, CA, USA

[1]http://sail.usc.edu, [2]www.isi.edu, [3]www.behavioralinformatix.com

Abstract

Automatically evaluating pronunciation quality of non-native speech has seen tremendous success in both research and commercial settings, with applications in L2 learning. In this paper, submitted for the INTERSPEECH 2015 Degree of Nativeness Sub-Challenge, this problem is posed under a challenging cross-corpora setting using speech data drawn from multiple speakers from a variety of language backgrounds (L1) reading different English sentences. Since the perception of non-nativeness is realized at the segmental and suprasegmental linguistic levels, we explore a number of acoustic cues at multiple time scales. We experiment with both data-driven and knowledge-inspired features that capture degree of nativeness from pauses in speech, speaking rate, rhythm/stress, and goodness of phone pronunciation. One promising finding is that highly accurate automated assessment can be attained using a small diverse set of intuitive and interpretable features. Performance is further boosted by smoothing scores across utterances from the same speaker; our best system significantly outperforms the challenge baseline.

Index Terms: Behavioral Signal Processing (BSP), computational paralinguistics, Goodness of Pronunciation (GOP), speech assessment, non-native speech, prosody, challenge

1. Introduction

Speech production is an intricate process involving multiple levels of planning and motor coordination in order to interweave segmental articulations and encode suprasegmental linguistic and paralinguistic information. Moreover, individuals differ in all facets of the speech production pipeline due to a variety of reasons (e.g., physical, environmental), leading to several sources of variability, including language background.

With the advancement of speech technologies, engineers have focused on creating assistive tools for language learning: from automatic literacy tutors [1], to the focus of this challenge, speech nativeness [2, 3]. Computer-Assisted Pronunciation Training (CAPT) is an invaluable resource for second-language (L2) learners that offers flexibility in scheduling at reduced costs. Languages differ in their phonemic, prosodic, and grammatical structures; quite often, L2 learners will retain certain speech attributes of their native language, leading to perceived abnormality, or “non-nativeness,” by L1 listeners. Our approach to this challenge is grounded in the automatic creation of informative pronunciation and prosodic features.

Prosody is at the core of effective human-human communication. It serves grammatical functions, such as segmenting utterances into phrases, pragmatic functions like differentiating statements from questions, and communicates attitude and affect. Since languages themselves have different rhythms and intonations, non-native speakers often use those prosodic characteristics of their first language, particularly when learning a stress-timed L2 given a syllable-timed L1, or vice-versa. Several recent studies have investigated L2 pronunciation through prosodic features. Levow et al. developed a pitch accent recognition algorithm for labeling non-native speeech which uses local and co-articulatory context [4]. Hansen et al. also proposed using source-generator based prosodic features to classify foreign accents of American English [5]; they showed that energy, duration, and spectral related features could play an important role in accent detection. Much research has been done on automatically separating native vs. non-native speakers [3, 6–8].

Phonemic identity and pronunciation quality are also important cues for automatically scoring degree of nativeness. Phoneme confusion between language makes both comprehension and production difficult for L2 learners. For instance, German speakers of English cannot pronounce /z/ clearly since there is no /z/ sound in the German language, while for Japanese, the lack of /r/ usually causes speakers to pronunce /r/ as /l/. Witt et al. [9] presented a goodness-of-pronunciation (GOP) measure to quantify phonetic pronunciation quality. They also improved their model by including the expected pronunciation errors in the recognition network, as did Black et al. [10] in the context of children’s literacy assessment.

Our approach is centered on combining knowledge-based and data-driven feature sets extracted at multiple time scales (segmental, suprasegmental). Much of these knowledge-inspired features are prosodic (suprasegmental) and pronunciation/articulatory (segmental) cues that were motivated and discussed in this section. Data-driven features have the advantage of fewer dependencies, with the idea being to discover trends through supervised learning methods; the downside is that they are naïve and suffer from high dimensionality. Conversely, knowledge-inspired features rely on other information that may be unreliable or noisy, but they are more intuitive and interpretable and have a lower dimensionality. Finally, we also consider unsupervised speaker clustering and smoothing methods, under the assumption that speakers will be perceived with consistent levels of nativeness across utterances.

2. Corpora & Baseline System

Four different corpora were analyzed: two make up the train set, and the other two are the development (“dev”) and test sets. Each corpus is comprised of multiple non-native speakers of English from a variety of language backgrounds (German, Italian, Chinese Japanese, French, Spanish, Hindi, other), although language ID was not provided. Each speaker read multiple sentences (disjoint across corpora). The degree of nativeness (*DN*) was labeled for each *utterance*, with higher scores representing higher degrees of *non-nativeness*. As shown in Fig. 1, the train and dev sets were rated on different scales; this was the primary reason for Spearman's correlation as the performance metric. *DN* scores and speaker IDs for the test set were not provided. Word-level transcriptions were available for each utterance in all data sets and included enriched markers for expected positions of major (B3) and minor (B2) phrase boundaries. The INTERSPEECH 2015 challenge organizers have requested that readers refer to the challenge paper for more details [12].

The *DN* Challenge baseline system consists of a purely data-driven approach: training a linear support vector regression (SVR) model on 6373 utterance-level static features by computing functionals of low-level descriptors (LLDs) that include prosodic (e.g., f_0, energy), voice quality (e.g., jitter, shimmer), and spectral (e.g., MFCCs, RASTA coefficients) cues. Please see [12, 13] for more details. We will compare the performance of this baseline system to our proposed methodology in Sec. 6.

3. Data Pre-processing

3.1. Forced Alignment & Voice Activity Detection

We exploited the available transcriptions by performing speech-text ("forced") alignment using freely available HTK [14] acoustic models (AMs) trained on out-of-domain native English speech [15]; we will leverage this "mismatch" in native vs. non-native speakers in Sec. 4.2. We used the CMU dictionary [16], with its multiple acceptable phonetic pronunciations for each word, and appended entries for any out-of-vocabulary words. From the output of the forced alignment, we extracted word, syllable, and phone boundaries (and the corresponding acoustic log-likelihoods). By using a grammar that allowed for the optional detection of inter-word pauses, this alignment process also acted as an accurate voice activity detector (VAD).

3.2. Speaker Clustering

As illustrated in Fig. 1, *DN* scores are correlated across utterances from the same speaker. Therefore, the identity of the speaker for each utterance can be utilized to smooth and improve *DN* score predictions. However, as described in Sec. 2, the speaker ID is unknown for the test utterances, so an unsupervised speaker clustering approach is needed. We used a bottom-up agglomerative hierarchical clustering (AHC) method [17] with k-means post-refinement [18]. Each cluster is modeled by a single Gaussian, and the generalized likelihood ratio [19] is used as the cluster distance metric [18]. Using the VAD (Sec. 3.1), we removed all periods of silence from the utterances before clustering. Since the number of speakers in the test set was known a priori (54, with 11 sentences per speaker), we stopped clustering once 54 distinct clusters were created.

4. Feature Extraction

4.1. Data-Driven Features

Our proposed data-driven features begin with the baseline LLDs extracted every 10ms with openSMILE [20]. Non-speech frames according to VAD (Sec. 3.1) are removed. In addition to computing utterance-level functionals as in the baseline system (Sec. 2), we also computed so-called "functionals-of-functionals," as proposed in [21] and used successfully in [22, 23]. LLD contours are first split into short disjoint windows of equal length L. Functionals are then computed *within* each window, and finally, functionals of these functionals are computed *across* all windows in the utterance. This process is meant to better capture moment-to-moment changes that occur at shorter time scales. In this work, we chose two different window lengths, $L \in \{0.1\text{s}, 0.5\text{s}\}$, based on the fact that more than 95% of the utterances in the corpora are more than one second in duration, which means the vast majority of utterances have a sufficient number of windows to compute reliable statistics. Because of the compounding nature of these calculations, there are 11323 data-driven functionals-of-functionals features in total.

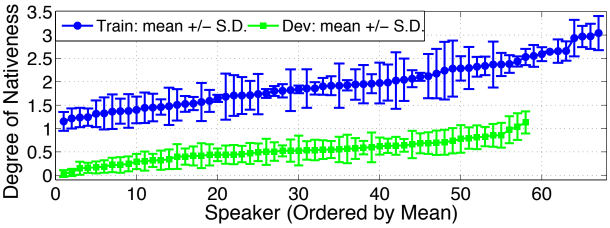

Figure 1: *Nativeness scores for speakers in the train/dev sets.*

4.2. Knowledge-Inspired Features

Prosody is the rhythm, stress, and intonation of speech. As such, we computed features that targeted these qualitative constructs. Immediately from the forced alignment (Sec. 3.1), we extracted important prosodic cues that captured *pausing* behaviors/strategies and *speaking rate*. Since the transcriptions had markers indicating phrase boundaries (Sec. 2), we first categorized each inter-word pause as one of the following: expected major (B3), expected minor (B2), or unexpected. Utterance-level pausing features were calculated for each combination of the 3 pausing categories in two ways: 1) percentage of the utterance duration due to pauses, and 2) mean duration of pauses.

We extracted two types of *speaking rate* features: rates (token / time) and durations (time / token). The former was calculated by creating a vector of rates ($1/\text{duration}$) for each token in the utterance and computing functionals (e.g., mean, S.D.) across it. We chose 3 different tokens at varying time scales: word, syllable, and phone. Similarly, the duration features were calculated by computing functionals of durations for different linguistic classes: syllables, phones, vowels, and consonants.

The next set of knowledge-driven features are related to lexical stress and speech *rhythm*. We compared the stress of each aligned phone to the stress labels in the CMU dictionary [16] that are provided for each vowel to signal data. For each aligned phone, we compute the mean f_0, energy, and intensity and then quantize these values into three quantiles for each utterance. Since the stress labels are also 3-valued, we define a distance measure based on exact-matching. For each signal (3), we obtain a measure that is the summation of mis-matches between the labeled and signal-derived stress.

We also quantify speech *rhythm* using pairwise variability indices (PVI, [24]) and global interval proportions (GIP, [25]). PVIs measure local changes in duration. We compute six PVI measures: normalized and unnormalized measures on consonant, vowel, and syllable durations. GIPs compute gross statistics on segmental durations; in particular, we included the percentage of vowel speech and S.D. of vowel and consonant durations within an utterance. Similar speech rhythm measures were previously considered for intoxicated speech detection [26].

Next, we implicitly model the stress and intonation of speech through *template*-based exemplar features, initially proposed for modeling children's prosody vs. an adult exemplar [27] and also successfully used in [28]. First, a single-prosodic functional for each token (phone or word) is calcu-

Type	N_{sig}/N	Best Feature: Spearman's Correlation		
		Description of Best Feature	*Train*	*Dev*
Pausing	7/8	Fraction unexpected pauses	**0.39**	**0.59**
Rate	18/47	Mean rate: phones/sec	**-0.43**	**-0.56**
Rhythm	5/14	Consonant PVI	**0.25**	**0.37**
Template	6/13	Phone duration mean |diff|	**0.37**	**0.49**
GOP	3/6	Utterance: phone+sil loop	**-0.48**	**-0.55**

Table 1: *Number of significantly correlated knowledge-inspired features ($p < 0.05$), along with the best performing feature.*

lated per feature: duration, median f_0, and median intensity. This is a time-aligned feature representation which we can compare with other readings. Template-based features are advantageous in that all computations are performed on the continuous-scale signal contours. However, since we do not have an exemplar production for each sentence, we must infer one from the other speakers. We take the mean feature contour from all other productions of the same sentence, assuming that this average will provide a suitable exemplar. In reference to this exemplar, we then computed 3 measures (Pearson's correlation, mean absolute difference, S.D.) for each of the following 4 contours: phone duration/pitch/intensity and word duration. Only sentences repeated 5 times were considered; otherwise, imputation with the mean value was used.

The last proposed knowledge-inspired feature is goodness-of-pronunciation (GOP) scoring, first introduced in [9] to detect phone-level pronunciation errors but also applied to L2 learning [29] and children's literacy assessment [10, 30]. The main idea behind GOP scoring is leveraging acoustic models (AMs) trained on native speakers to quantify pronunciation quality:

$$\text{GOP} = \frac{1}{N} \log \left(\frac{P(O|\text{Transcript})}{P(O|\text{AM loop})} \right) \tag{1}$$

, where N is the number of frames and O are the acoustic features (MFCCs in this case). The numerator is the likelihood of the acoustics, given the transcription and the native AMs, which is equivalent to forced alignment (Sec. 3.1). The denominator is estimated via automatic speech recognition with the same set of AMs and an "AM loop" grammar. Higher GOP scores indicate higher pronunciation quality, whereas lower GOP scores suggest the speech is a poor match to the native AMs. In this work, we experimented with 3 different AM loops (phones+silence, phones only, silence only) and computed GOP scores over different temporal regions: full utterance (including inter-word pauses), speech regions only, vowels only (based on [11]).

Table 1 shows that almost half of the knowledge-based features are significantly correlated with the *DN* scores in the train and dev sets ($p < 0.05$). We restrict our analysis to the best performing feature for each proposed type due to space constraints. As shown in Table 1, speakers sound more *non-native* when they: pause more unexpectedly; speak at a slower average rate; have higher local variability in consonant durations; differ more than other productions of the same sentence; have lower GOP scores. All of these findings agree with intuition.

5. Predicting Degree of Nativeness

We experimented with several supervised machine learning techniques to map the various features to the *DN* scores, including regression and ranking algorithms [31], ensemble methods (e.g., bagging, boosting [32], stacking [33]), and early vs. late fusion. Due to space contraints, we only describe our best performing system, shown in Fig. 2. For parameter-tuning purposes and to prevent overfitting, the train set was split into 6 approximately equal-sized speaker-disjoint cross-validation folds.

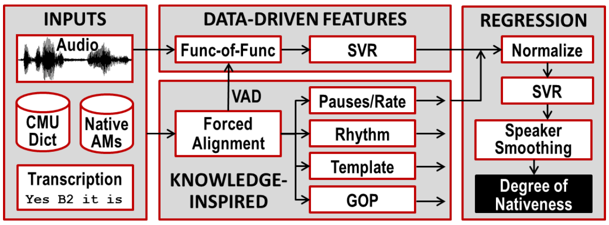

Figure 2: *Automatically evaluating degree of nativeness using support vector regression (SVR), trained on data-driven and knowledge-inspired features, with speaker-level smoothing.*

5.1. Support Vector Regression & Fusion

For the data-driven functionals-of-functionals ("FoF"), since dimensionality is so high, we first employed unsupervised feature reduction. A feature was dropped unless the following two criteria were both met for all cross-validation folds and datasets: 1) the feature must have a coefficient of variation ≥ 0.01 (to eliminate *irrelevant* features), and 2) the feature must *not* have a Pearson's or Spearman's correlation ≥ 0.98 with any other feature (to eliminate *redundant* features). When two features were highly correlated, we dropped the one that was less normally distributed, based on the Kolmogorov-Smirnov test. This process reduced the dimensionality by 34% (from 11323 to 7478).

The feature matrix was then normalized/scaled four different ways: Z-normalization, i.e., subtracting the mean and dividing by the S.D., 1) across all data; 2) for each fold and dataset separately; and scaling the feature matrix to be bounded between 0 and 1, i.e., subtracting the minimum and dividing by the range, 3) across all data; 4) for each fold and dataset separately. Finally, we trained l_1-loss and l_2-loss l_2-regularized linear support vector regression (SVR) models, as implemented in LIBLINEAR [34], with the loss functions defined as:

$$\min_{\boldsymbol{w}} \left(\boldsymbol{w}^T \boldsymbol{w} + C \sum_{u} \max \left(0, \left| DN_u - \boldsymbol{w}^T \boldsymbol{x}_u \right| - \epsilon \right)^l \right), l \in \{1,2\} \tag{2}$$

, where $\boldsymbol{x}_u$ is the feature vector for utterance u, $\boldsymbol{w}$ is the linear weight vector, l determines the type of loss function (l_1 or l_2), ϵ is the loss sensitivity parameter, and C is the regularization parameter. A grid search was used to tune ϵ and C. The "optimal" parameter values, normalization technique, and SVR loss type, i.e., the combination that attained the highest mean Spearman's correlation across the 6 train folds, was then applied to the dev set after retraining the SVR model on the full train set.

Feature-level ("early") fusion of the proposed knowledge-inspired features was possible because of their low dimensionalities. Fusion between the data-driven and knowledge-inspired features was attained by treating the output prediction scores of the data-driven FoF SVR model as one additional "feature" (Fig. 2). After removing irrelevant/redundant features and normalizing/scaling the feature matrix as before, we trained similar l_1-loss and l_2-loss linear SVR models on these fused features.

5.2. Speaker-level Smoothing

As motivated in Sec. 3.2 and shown in Fig. 1, *DN* scores are correlated across a speaker's utterances, suggesting that directly modeling this dependency may improve prediction performance. We experimented with the following smoothing tech-

Category	Feature	N	Train		Dev
			mean	*S.D.*	
Baseline	[12]	6373	**0.40**	—	**0.42**
Time Scale	Utterance	530	**0.37**	0.12	**0.29**
Time Scale	FoF (0.1s)	3100	**0.48**	0.08	**0.44**
Time Scale	FoF (0.5s)	3848	**0.47**	0.06	**0.38**
All Proposed	All Proposed	7478	**0.48**	0.06	**0.45**

Table 2: *Dimensionality and performance (Spearman's correlation) of the support vector regression model for various subsets of data-driven features on the train set (6 folds) and dev set.*

nique in (3), where $\widehat{DN}_u^s$ is the predicted DN score of utterance u from speaker s, and w_u^s is the number of syllables in u:

$$\widetilde{DN}_u^s = (1-\alpha)\widehat{DN}_u^s + \frac{\alpha}{\sum\limits_{u\in s} w_u^s} \sum_{u\in s} w_u^s \widehat{DN}_u^s, \quad 0 \leq \alpha \leq 1 \tag{3}$$

Therefore, $\widetilde{DN}_u^s$ is a smoothed linear combination of the utterance's predicted score $\widehat{DN}_u^s$ with the speaker's *weighted average* score (with longer utterances containing more syllables carrying more weight when computing this average). The decision to use this weighted average was made since there is more acoustic evidence in longer utterances, giving human raters and computational methods alike more of a chance to make a reliable decision. In (3), tuning parameter α determines the level of smoothing; $\alpha = 0$ means no smoothing, i.e., the utterance's predicted score is applied, while $\alpha = 1$ means the speaker's average score is applied to all of the speaker's utterances.

6. Results & Discussion

In this section, we analyze the performance of our various data-driven and knowledge-inspired sub-systems. First, we examine performance of SVR using data-driven functional-of-functional (FoF) features on the train and development sets (Table 2). Our reduced set of utterance-level functionals (i.e., 530 vs. 6373) show a drop in performance relative to the baseline. However, through modeling these features at various time scales, FoF features exceed baseline performance on train and dev; in both settings, FoF at 0.1s time-scale is the top performing individual feature set. We note that as the number of features increases, the standard deviation of model performance across training folds decreases, an indication of a consistent linear-kernel SVR model. Fusion of all proposed FoF features exceeds baseline performance—evidencing the benefits of the rigorous temporal modeling in this framework.

Knowledge-inspired feature performance is shown in Table 3, including score-level fusion with data-driven FoF features. Speaking rate features perform better than any other feature group on both train and dev, followed closely by pausing features; this reflects the critical importance of timing to perceived nativeness. GOP features alone (6 features) exceed baseline performance, indicating that articulation cues can identify non-native speech. Rhythm and template features also have significant performance, albeit below the other feature groups. Fusion of all knowledge-inspired features leads to performances well-above the baseline (listed as LOO: Func-of-Func). Leave-one-out experiments do not suggest that any feature group greatly degrades performance; on the contrary, fusion of all proposed features achieves peak performance of 0.605, 0.707, and 0.701 on the train, dev, and test sets, respectively.

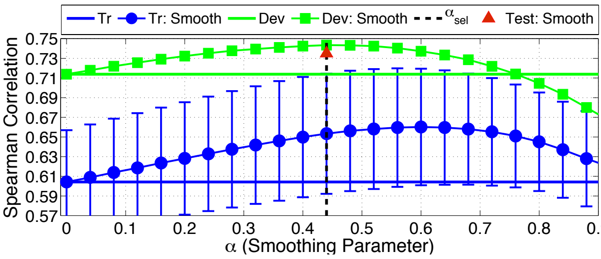

Figure 3: *Improving utterance-level score prediction by smoothing with the speaker's average score across all utterances.*

Feature/Classifier	N	Train		Dev	Test
		mean	*S.D.*		
Baseline [12]	6373	**0.403**	—	**0.415**	**0.425**
Func-of-Func (FoF)	7478	**0.483**	*0.065*	**0.454**	—
Pausing	8	0.405	*0.107*	0.613	—
Speaking Rate	47	**0.552**	*0.047*	**0.627**	—
Rhythm	14	0.262	*0.048*	0.411	—
Template	13	0.370	*0.104*	0.455	—
GOP	6	0.430	*0.066*	0.444	—
LOO: Func-of-Func	88	**0.585**	*0.055*	**0.690**	—
LOO: Pausing	80+1	0.603	*0.051*	0.700	—
LOO: Speaking Rate	41+1	0.584	*0.047*	**0.701**	—
LOO: Rhythm	74+1	0.603	*0.048*	0.703	—
LOO: Template	75+1	**0.604**	*0.052*	0.688	—
LOO: GOP	82+1	0.596	*0.053*	0.699	—
All Proposed	88+1	**0.605**	*0.050*	**0.707**	**0.710**
Spkr Smooth; α=0.44	88+1	**0.653**	*0.061*	**0.744**	**0.735**

Table 3: *Dimensionality and performance (Spearman's correlation) for each of the proposed features on the train set (6 folds) and dev/test sets. "+1" represents score-level fusion with the Func-of-Func (FoF) classifier. "LOO" means Leave-One-Out.*

Lastly, we perform speaker-level smoothing of our best SVR model, utilizing the assumption that non-native speakers are consistently non-native. Optimization is shown in Figure 3, where optimal performance is found for $\alpha = 0.44$. Smoothing provides a moderate improvement to 0.744 on the dev set; test set performance with this system is relatively consistent at 0.735, significantly outperforming the baseline correlation of 0.425 ($p < 0.001$).

7. Conclusions & Future Work

In this paper, we showed that a combination of low-dimensional knowledge-inspired segmental and suprasegmental features (pausing, speaking rate, rhythm, and goodness-of-pronunciation) are able to accurately match subjective Degree of Nativeness. In combination with a data-driven feature system and speaker-level smoothing, we achieved highly accurate evaluation of non-native pronunciation quality.

Future work will include experimentation with additional fusion strategies as well as novel ways to incorporate a speaker's L1 (i.e., exploiting specific non-native trends that are more prone to occur). Given that English is a stress-timed language, it may be beneficial to utilize metrics that measure variability between stressed syllables, rather than all adjacent syllables as we have currently done (i.e., isochrony features [35]).

8. References

[1] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, Nov 2003, pp. 186–191.

[2] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.

[3] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for english as L2," *Proc. Speech Prosody, Chicago*, 2010.

[4] G.-A. Levow, "Investigating pitch accent recognition in non-native speech," in *Proceedings of the ACL-IJCNLP Conference Short Papers*, 2009, pp. 269–272.

[5] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *IEEE ICASSP*, 1995, pp. 836–839.

[6] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters," in *INTERSPEECH*, 2008.

[7] F. Hnig, A. Batliner, K. Weilhammer, and E. Nth, "Islands of failure: Employing word accent information for pronunciation quality assessment of english l2 learners," in *Proceedings of SLATE, Wroxall Abbey*, 2009.

[8] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for ted talks," in *IEEE ICASSP*, 2011, pp. 5672–5675.

[9] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[10] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1015–1028, 2011.

[11] L. Chen, K. Evanini, and X. Sun, "Assessment of non-native speech using vowel space characteristics," in *Spoken Language Technology Workshop*, 2010, pp. 139–144.

[12] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in *INTERSPEECH*, 2015.

[13] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.

[14] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," Cambridge University Engineering Department, Tech. Rep., Dec. 2006.

[15] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cavendish Laboratory, Tech. Rep., 2006.

[16] R. L. Weide, "CMU pronouncing dictionary," Carnegie Mellon University, 1994. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict/

[17] W. Wang, P. Lv, and Y. H. Yan, "An improved hierarchical speaker clustering," *Acta Acustica*, 2008.

[18] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, Nov 2008.

[19] A. S. Willsky and H. L. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 108–112, Feb 1976.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[21] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *IEEE ICASSP*, Las Vegas, NV, USA, 2008, pp. 4501–4504.

[22] M. P. Black, A. Katsamanis, B. Baucom, C.-C. Lee, A. Lammert, . A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features" *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[23] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors" *Computer Speech & Language*, vol. 28, no. 2, pp. 375–391, 2014.

[24] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[25] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," 2002.

[26] F. Hönig, A. Batliner, and E. Nöth, "Does it groove or does it stumble-automatic classification of alcoholic intoxication using prosodic features." in *INTERSPEECH*, 2011, pp. 3225–3228.

[27] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, p. 14, 2011.

[28] D. Bone, T. Chaspari, K. Audhkhasi, J. Gibson, A. Tsiartas, M. Van Segbroeck, M. Li, S. Lee, and S. S. Narayanan, "Classifying language-related developmental disorders from speech cues: the promise and the potential confounds." in *INTERSPEECH*, 2013, pp. 182–186.

[29] A. Neri, C. Cucchiarini, and H. Strik, "The effectiveness of computer-based speech corrective feedback for improving segmental quality in l2 dutch," *ReCALL*, vol. 20, no. 2, pp. 225–243, May 2008. [Online]. Available: http://dx.doi.org/10.1017/S0958344008000724

[30] J. Tepperman, M. P. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. S. Narayanan, "A bayesian network classifier for word-level reading assessment," in *Proceedings of InterSpeech*, Antwerp, Belgium, Aug. 2007, p. 21852188.

[31] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *The Journal of machine learning research*, vol. 4, pp. 933–969, 2003.

[32] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[33] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, vol. 54, no. 3, pp. 255–273, 2004.

[34] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[35] I. Lehiste, "Isochrony reconsidered." *Journal of phonetics*, 1977.