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

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

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

\textsuperscript{1}http://sail.usc.edu, \textsuperscript{2}www.isi.edu, \textsuperscript{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 speech 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 /h/ usually causes speakers to pronounce /h/ 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., f0, 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 ∈ {0.1s, 0.5s}, 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.

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 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/duration) for each token in the utterance and computing functionals (e.g., mean, S.D.) across it. We chose 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 f0, energy, and intensity and then quantize these values into three quantities 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-
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 constraints, 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.

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:

$$\min_w \left( w^T w + C \sum_u \max(0, DN_u - w^T x_u - \epsilon) \right), \ell \in \{1, 2\}$$

where $x_u$ is the feature vector for utterance $u$, $w$ is the linear weight vector, $l$ determines the type of loss function ($l_1$ or $l_2$), $\epsilon$ is the loss sensitivity parameter, and $C$ is the regularization parameter. A grid search was used to tune $\epsilon$ 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 dimensionality. 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, suggesting that directly modeling this dependency may improve prediction performance. We experimented with the following smoothing tech-
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 of nativeness. In combination with a data-driven feature system and speaker-level smoothing, we achieved highly accurate evaluation of non-native pronunciation quality.

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


