Automatic Assessment of Non-Native Prosody for English as L2

Florian Hönig[1], *Anton Batliner*[1], *Karl Weilhammer*[2], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, Universität Erlangen-Nürnberg, Germany
[2]digital publishing, München, Germany
{hoenig,batliner}@informatik.uni-erlangen.de

Abstract

We recorded non-native English productions of 55 speakers; a subset of these productions was assessed by 60 native English speakers as for their quality w. r. t. intelligibility, rhythm, etc. Applying multiple linear regression on a large prosodic feature vector – modelling approaches known from the literature as well as generic prosody – we can automatically predict the listener's assessments with correlations of up to .85. We discuss most important features and limitations of this approach.

Index Terms: non-native prosody, rhythm, intelligibility, foreign accent, linear correlation

1. Introduction

Not only segmental errors but suprasegmental 'peculiarities' can impede the understanding of L2 learners' productions and have to be modelled in computer-assisted pronunciation training (CAPT) [1, 2]. Such prosodic phenomena, located on the word level and above, encompass word accent position, sentence prosody, and rhythm, and help listeners to structure the speech signal and to process segmental, syntactic, and semantic content successfully. A few studies deal with non-native accent identification using prosodic parameters [3, 4]. In [5], the automatic detection of erroneous word accent positions in English as L2 is addressed. Suprasegmental native traits have been, e. g. investigated recently in basic research when trying to model language-specific rhythm [6, 7]. Maybe the most important general factor to be modelled in CAPT is non-native rhythm: the English prosody of, e. g. French, Spanish, or Hindi native speakers can sound 'strange'. The reason is a difference in rhythm that has been noted amongst others by [8], p. 97, who speaks about syllable timed languages such as French ("the syllables [...] recur at equal intervals of time – they are *isochronous*"), and stress-timed languages such as English ("the stressed syllables [...] are isochronous"). [6] and [7] challenge this traditional terminology because in empirical studies, such an isochrony could not be observed; they claim that it is rather a more complicated constellation where especially syllables not carrying the word accent, that are weak (schwa) in 'stress-timed' languages, are produced stronger in 'syllable-timed' languages. Thus we might expect such differences to show up in L2 learners whose native language L1 does not display the native structure of L2.

In the following, we first describe in Sec. 2 our database with English L2 productions of speakers with different L1, and the perception experiment carried out with native English speakers who had to judge these productions. In Sec. 3 we then present the features used for the assessment experiments dealt with in Sec. 4; limitations are addressed in Sec. 5.

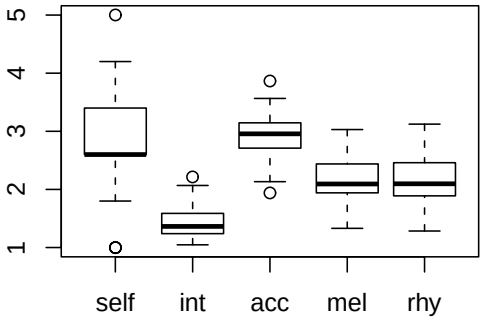

Figure 1: Box-and-whisker plot of the the speaker-level scores (averaged across all 60 labellers)

2. Material and human assessment

We recorded 55 English L2 speakers: 25 German, 10 French, 10 Spanish, and 10 Italian speakers, and additionally four native American English (AE) 'reference' speakers. They had to read aloud 329 utterances shown on the screen display of an automated recording software, and were allowed to repeat their production in case of false starts etc. Only the last token, i. e. the one supposed to be error-free — or at least as good as possible, was taken for further processing. The data to be recorded consisted of two short stories (broken down into sentences to be displayed on the screen), sentences containing, amongst other, different types of phenomena such as intonation or position of phrase accent (*This is a house.* vs. *Is this really a house?*), or tongue-twisters, and words/phrases such as *Arabic/Arabia/The Arab World/In Saudi-Arabia, ...*; pairs such as *'subject* vs. *sub'ject* had to be repeated after the prerecorded production of a tutor. Some sentences were taken from the ISLE corpus [9].

Based on annotations of three experienced labellers [5], we defined the subset of the following five sentences that were judged as 'prosodically most error-prone for L2 speakers of English':

> *We're planning to travel to Egypt for a week or so.*
> *Can I have soup, then lamb with boiled potatoes, green beans and a glass of red wine?*
> *They will have to transport the components overland.*
> *The referee needed a police escort after the match.*
> *The company expects to increase its workforce next year.*

The perception experiment was conducted web-based, using the tool PEAKS [10]. 20 native AE, 19 native British English (BE), and 21 native Scottish English (SE) speakers with normal hearing abilities judged each sentence in random order. The questions were:

1. DID YOU UNDERSTAND WHAT THE SPEAKER SAID?
   (1) *yes, the sentence is completely understandable* (2) *yes, but*

Table 1: *Correlations between the speaker-level scores of labellers and groups of labellers. The column headings refer to the sets of labellers that are compared. 1A/RA: average correlation of **1** **A**merican labellers' scores with the averaged scores from the **R**emaining **A**merican labellers; HA/HA: average correlation of **H**alf of the **A**merican Labellers with the other half; similar for the **B**ritish labellers, (1B/RB and HB/HB) and the **S**cottish (1S/RS and HS/HS). 1/R refers to the comparison of **1** (American, British or Scottish) labeller with all **R**emaining labellers. AA/AB compares **A**ll **A**merican with **A**ll **B**ritish, AA/AS **A**ll **A**merican with **A**ll **S**cottish, and AB/AS **A**ll **B**ritish with **A**ll **S**cottish labellers. H/H compares **H**alf of all (American, British or Scottish) labellers with the remaining half.*

	1A/RA	HA/HA	1B/RB	HB/HB	1S/RS	HS/HS	1/R	AA/AB	AA/AS	AB/AS	H/H
int	0.41	0.77	0.53	0.85	0.45	0.83	0.50	0.95	0.96	0.97	0.92
acc	0.53	0.91	0.58	0.91	0.70	0.93	0.61	0.94	0.96	0.96	0.98
mel	0.35	0.83	0.43	0.77	0.50	0.79	0.45	0.97	0.96	0.97	0.96
rhy	0.37	0.85	0.56	0.85	0.60	0.88	0.52	0.96	0.97	0.96	0.98

some parts are not easy to understand (3) *well, the sentence needs some effort to be understood* (4) *no, most parts of the sentence are not easy to understand* (5) *no, the sentence cannot be understood at all*

2. DID YOU HEAR A FOREIGN, NON-ENGLISH ACCENT?
(1) *no* (2) *very slight* (3) *some accent* (4) *strong accent* (5) *extreme accent*

3. WHAT KIND OF ACCENT DO YOU THINK THIS SPEAKER HAS?
(1) *American* (2) *British* (3) *French* (4) *Italian* (5) *German* (6) *Spanish* (7) *Russian* (8) *Japanese* (9) *Indian* (10) *I don't know*

4. THIS SENTENCE'S MELODY SOUNDS...
(1) *normal* (2) *acceptable, but not perfectly normal* (3) *slightly unusual* (4) *unusual* (5) *very unusual*

5. THE ENGLISH LANGUAGE HAS A CHARACTERISTIC RHYTHM (TIMING OF THE SYLLABLES). HOW DO YOU ASSESS THE RHYTHM OF THIS SENTENCE?
(1) *normal* (2) *acceptable, but not perfectly normal* (3) *slightly unusual* (4) *unusual* (5) *very unusual*

For the five selected sentences spoken by 55 speakers, we collected a total of 82500 labels from 60 labellers. For the present study, we concentrate on the labels for intelligibility (*int*), accent (*acc*), melody (*mel*) and rhythm (*rhy*), i. e. the answers to questions 1, 2, 4 and 5. For each speaker, the labels on the Likert scales were averaged over all five sentences to get a single score for each criterion. Figure 1 shows a box-and-whisker plot of the resulting scores; in addition, *self* represents the speakers' self-assessment, based on a mapping of CEF levels[1] (A1 to C2) onto a corresponding scale between 1 and 5. We see that *self* is close to *acc*, we can speculate that the difference between *acc* and *mel*/*rhy* is due to additional segmental errors, and that the lower (i. e., better) *int* value might be traced back to the listeners' language model which is not fully impaired by segmental or suprasegmental errors. Table 1 shows inter-labeller agreements (Pearson correlation coefficient *r*); where scores are taken from a group of labellers, they are computed as the arithmetic mean of each of the labeller's scores. The scores from a single labeller correlate moderately with the averaged scores from the other labellers (e. g. $0.35 \leq r \leq 0.53$ when comparing one of the AE labellers with the other AE labellers, column "1A/RA"). The averaged scores from multiple labellers, however, show higher correlations (e. g. $r \geq 0.77$ when comparing half of the American labellers with the other half, column "HA/HA"). Also the averaged AE, BE, and SE scores correlate highly ($r > 0.94$, columns AA/AB to AB/AS). From Table 2, which shows *r* for the different scoring criteria (averaged over all labellers), it turns out that the ratings are highly correlated among themselves. As *mel* and *rhy* highly correlate with each other (0.98), we derive the combined supra-segmental measure *sup* which is computed by normalising and averaging these two measures.

[1] www.cambridge.org/catalogue/catalogue.asp?isbn=9780521005319

3. Features

We compute a large number of features measuring different prosodic traits. All processing is done fully automatic; however, we assume that the spoken word sequence is identical with the utterance the speaker had to read. For each utterance, the DC is removed and the maximal amplitude of the signal is normalized. Word, syllable, and phoneme boundaries are obtained from a forced alignment using cross-word triphon HMMs trained on native AE. Short-time energy and fundamental frequency (F0) are computed on a frame-by-frame basis (step size 10 msec). F0 is logarithmised and normalized, and interpolated during non-voiced segments. Syllable nuclei are defined as the vocalic part of syllables. A syllable is considered as stressed if it is a mono-syllabic word bearing a primary or secondary phrase accent – this decision has been based on the productions of our 4 native reference speakers – or if it is part of a multi-syllabic word, and having either secondary or primary word accent.

Table 2: *Correlations of the speaker-level scores (averaged across all labellers) amongst each other.*

	int	acc	mel	rhy	sup
int	1.00				
acc	0.88	1.00			
mel	0.85	0.86	1.00		
rhy	0.89	0.89	0.98	1.00	
sup	0.87	0.88	0.99	0.99	1.00

3.1. Specialized Prosodic Feature Sets

Speech Rate Measures: We extracted the rate of stressed and unstressed syllables, and the rate of vocalic segments for each sentence (the rates are defined as number of units per second), and derive six speaker-level features for the speech rate (*SR*) by calculating the mean and standard deviation of each of these three sentence-level features.

Isochrony Features: In order to capture possible isochrony properties [8], we calculate the distances between the centers of consecutive stressed and unstressed syllables. The centers are identified as the frames with maximal short-time energy within the nuclei. We compute six sentence-level features: mean distances between stressed and between unstressed syllables, standard deviations of those distances, and the ratio of means and standard deviations. 12 speaker-level features *Iso* are calculated from mean and standard deviation of each of these sentence-level features.

Variability Indices: Following [6], we identify vocalic and consonantal segments and calculate the raw Pairwise Variability Index (rPVI) which is defined as the absolute difference in du-

Table 3: *Prosodic features and their context. Bullets (filled circles) indicate which contexts in columns 2–6 are used for the 31 local features; for the 100 context-capturing features, additionally the contexts indicated by empty circles are used. Curly brackets indicate that all the features displayed in these three rows are computed for all contexts in the three rows in columns 2–6.*

features for the actual unit '0' computed from up to ± units	**context size** -2	-1	0	1	2
Dur: Norm, Abs; En: RegCoeff, MseReg, Mean, Abs, Norm; F0: RegCoeff, MseReg, Mean		○	●	○	
	○		●		
		○		○	
En: Max, MaxPos		○	●	○	
F0: Max, MaxPos, Min, MinPos		○	●	○	
F0: Off, Offpos		○	●		
F0: On, Onpos			●	○	
Pause-before		○	○		
Pause-after			●	○	

ration of consecutive segments and its normalized version nPVI (rPVI divided by the mean duration of the segments) for vocalic and consonantal segments. From these four utterance-level features, we compute 8 speaker-level Pairwise Variability Index features *PVI* (mean and standard deviation of the sentence-level features).

Global Interval Proportions: Following [7], we compute the percentage of vocalic intervals (of the total duration of vocalic and consonantal segments), and the standard deviation of the duration of vocalic and consonantal segments of a sentence. Again, by computing mean and standard deviation of these sentence-level features, we derive six speaker-level features measuring Global Proportions of Intervals (*GPI*).

3.2. General-Purpose Prosodic Features

In addition to the specialized features, we apply our comprehensive general-purpose prosody module which has already been successfully applied to as diverse problems as phrase accent and phrase boundary recognition [11], word accent position classification [5], and emotion recognition [12]. The features are based on duration, energy, pitch, and pauses, and can be applied to arbitrary units of speech (here, the units used are words, syllables, and nuclei). Some of the energy and duration based features are normalized versions of a quantity, e. g. the duration of a word divided by the average duration of that specific word. The statistics necessary for these normalization measures have been estimated on the native AE reference speakers' data.

Trying to be as exhaustive as possible, we use a highly redundant feature set leaving it to data-driven methods to find out the relevant features. However, the procedure is based on knowledge and not on brute force. Features are extracted from a context of one or two units. A couple of neighbouring units are used to extract features for the current unit: For a local variant, the current unit and the context including the current and the following unit are used for feature extraction. A context-capturing variant uses contexts up to ± 2 neighbouring units. This process is detailed in Table 3. The features are abbreviated as follows: *duration features 'Dur':* absolute (Abs) and normalised (Norm); *energy features 'En':* regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalised (Norm) values; *F0 features 'F0':* regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis; *length of pauses 'Pause':* silent pause before (Pause-before) and after (Pause-after).

Additionally, we compute four features over a larger context of ± 7 units (or less, if the utterance is shorter), representing local estimates of global properties: *RateOfSpeech* (average speech rate), *DurTauLoc* (average duration), *EnTauLoc* (average energy) and *F0MeanGlob* (average fundamental frequency). These are appended to both the local and context-capturing configuration, ending up with 35 and 104 features per unit, respectively. A more detailed overview of the prosodic features is given in [11].

We now use all these prosodic features computed over different units and contexts to construct (again highly redundant) extensions of the *Iso*, *PVI* and *GP* features (in total 523): **(1)** we compute the context-capturing features for all stressed syllables and nuclei of stressed syllables, and the local features for all words, syllables and nuclei, and use the mean values of these features as speaker-level *extended Iso* features (2*104 + 3*35 = 313 features); words, syllables and nuclei at the start and end of the sentences that do not provide enough context units are skipped; **(2)** for the *extended PVI* features, the mean absolute difference of the local features of consecutive words, syllables and nuclei (3*35 = 105 features) is computed; **(3)** the standard deviations of the local features of all words, syllables and nuclei (3*35 = 105 features) represent the *extended GPI* features. These extended specific *Iso*, *PVI*, and *GPI* features are pooled into the general-purpose prosodic feature vector *Pros*.

3.3. Speech Recognition Features

We run a speech recognizer with a unigram language model on each sentence and compute word accuracy and word correctness for the resulting hypothesized word sequence. Additionally, we calculate the ratio of the lengths of the hypothesized word sequence and of the reference word sequence (productions of the native reference speakers). From these three sentence-level features, we compute mean and standard deviation yielding six speaker-level features *WR*.

4. Experiments and Results

We apply cross-validated forward feature selection (CV-FS) and simple multiple linear regression to predict the speaker level labels. For different setups of features and target labels, we evaluate the performance of the regression in terms of the average Pearson *r* in a 10-fold crossvalidation (CV). For the feature selection (which is performed in each CV fold), we use a wrapper approach and iteratively add the feature that leads to the best performance, which is estimated as the average performance of multiple linear regression on the current training set in a (nested) 10-fold CV. This process is continued until performance drops or the number of selected features exceeds 5.

Table 4: *Results on speaker level, after CV feature selection (CV-FS) to 5 features*

Scale	**Feature Types** SR	Iso	PVI	GPI	Pros	WR	All
int	0.46	0.63	0.45	0.42	0.70	0.72	0.76
acc	0.56	0.54	0.44	0.39	0.67	0.72	0.78
mel	0.74	0.70	0.51	0.56	0.76	0.55	0.81
rhy	0.71	0.72	0.56	0.52	0.83	0.64	0.85
sup	0.73	0.73	0.54	0.55	0.76	0.61	0.85

For space reasons, Table 4 is based only on all 60 labellers taken together, leaving aside differences between AE, BE, and SE; the overall performance of the different types of features is straightforward: amongst the specialized feature types, both *SR* and *Iso* are better than both *PVI* and *GPI*; *Pros* is better than *SR* and *Iso*. *WR* outperforms *Pros* as for *int* and *acc* which both are not only suprasegmental phenomena. In contrast, *Pros* outperforms *WR* if it is about the prosodic phenomena *mel*, *rhy*, and thus of course *sup*. Interestingly, the combination of prosodic features with *WR* features always yields improved performance, between 0.02 and 0.11, cf. All. Here, only *Pros* and *WR* features survive the CV-FS to 5 features, and there is always the same *WR* feature (mean word accuracy) amongst the 5 best. (The set of selected features can differ for each CV fold. We interpret the 5 features that are selected in a CV-FS on the whole training set. They are similar to the features selected in the individual CV folds.)

Again for space reasons, we confine our detailed interpretation of the 'surviving' 5 features to *sup* as the most interesting generic phenomenon from a prosody point of view; we have to keep in mind that this is a multi-variate analysis, thus results will be different if we assess the impact of single features. To make the regression coefficients $\beta_1, \beta_2, \ldots, \beta_5$ independent of the scale of the input, the variance of the features is standardized to one. The five features nicely comprise duration, energy, pitch, and word recognition:

1. *the mean (non-normalized) duration of all units consisting of two consecutive syllable nuclei* ($\beta_1 = 0.61$): the longer this combined measure is, the less native is the production; this is most likely some speech rate measure, assessing reading performance as well: non-native L2 speakers produce longer units due to hesitations, etc.
2. *the mean word accuracy* ($\beta_2 = -0.33$): this is simply word recognition: bad word recognition goes along with non-native traits.
3. *the standard deviation of the normalized word duration* ($\beta_3 = 0.21$): this is a global variability measure: it measures how much the word duration deviates from native productions.
4. *the mean normalized energy of the two syllable nuclei following stressed syllables* ($\beta_4 = 0.16$): in our sentences, often two unstressed syllables follow a stressed one; thus energy higher than the one produced by native speakers in these contexts indicate non-native productions.
5. *the mean of the mean square error of the F0 regression line for those syllable nuclei that follow a stressed syllable* ($\beta_5 = 0.16$): a high F0 variability in unstressed syllables following a stressed one indicates non-nativeness.

5. Discussion and Concluding Remarks

Obviously, the labellers could not reliably tell apart *mel* from *rhy*. Note that the design of this experiment was rather straightforward: the order of questions remained identical throughout; however, we doubt that a more sophisticated design would have changed matters much. The missing discrimination might be partly due to the specific speaking style 'reading aloud rather complex non-native sentences': this is a rather special style which does not really display something like 'integrating prosody' which might be a pre-requisite of native or non-native intonation to show up. Therefore, we as well cannot really say whether our results can be extrapolated to native production, i. e. whether the specialized feature sets really are inferior to extended specialized features taken from our general-purpose features. Note, moreover, that almost half of our speakers (25 out of 55) are L1 speakers of German which is 'stress-timed'; this fact might have 'smeared' any tendency towards non-native 'syllable-timed' traits in English as L2. Up to a large extent, we might in fact really assess reading capabilities (in a foreign language). Note, however, that this situation is common for almost all training programs. And of course, reading competence and linguistic/phonetic competence do have much in common. However, it seems plausible that the underlying phenomenon – which is modelled in different aspects of the specialized prosodic feature sets – can be modelled more accurately if we enrich the specialized features (*SR, Iso, PVI, GPI*) with information on prosodic content, such as duration, energy, or pitch, entailed in our extended specialized features, taken from general-purpose features. Using this strategy, we obtained a high correlation of 0.85 for the automatic assessment of prosody.

6. Acknowledgements

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the project *C-AuDiT - Computerunterstütztes Aussprache- und Dialog-Training* under Grant 01IS07014B. The responsibility lies with the authors. The perception experiments were conducted by Susanne Burger (Pittsburgh) and Catherine Dickie (Edinburgh). We want to thank Andreas Maier for adapting PEAKS to our task.

7. References

[1] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, Univ. of Cambridge, 1999.

[2] C. Cucchiarini, A. Neri, F. de Wet, and H. Strik, "ASR-based pronunciation training: Scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners," in *Proc. Interspeech*, Antwerp, 2007, pp. 2181–2184.

[3] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters," in *Proc. Interspeech*, Brisbane, 2008, pp. 759–762.

[4] J. Tepperman and S. Narayanan, "Better Nonnative Intonation Scores through Prosodic Theory," in *Proc. Interspeech*, Brisbane, 2008, pp. 1813–1816.

[5] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners," in *Proceedings of SLATE*, Wroxall Abbey, 2009.

[6] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 515–546.

[7] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proc. Speech Prosody*, Aix-en-Provence, 2002, pp. 115–120.

[8] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: University Press, 1967.

[9] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proc. LREC*, Athens, 2000, pp. 957–964.

[10] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS - A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, pp. 425–437, 2009.

[11] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The Prosody Module," in *Verbmobil: Foundations of Speech-to-Speech Translations*, W. Wahlster, Ed. Berlin: Springer, 2000, pp. 106–121.

[12] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of Tuning – Prototyping for Automatic Classification of Emotional User States," in *Proc. Interspeech*, Lisbon, 2005, pp. 489–492.