Assessing the Prosody of Non-Native Speakers of English: Measures and Feature Sets

Eduardo Coutinho [1,2], Florian Hönig [3], Yue Zhang [1], Simone Hantke [4]
Anton Batliner [3,5], Elmar Nöth [3], Björn Schuller [1,5]

[1]Department of Computing, Imperial College London, UK
[2]Department of Music, University of Liverpool, UK
[3]Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nuremberg, Germany
[4]Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
[5]Chair of Complex & Intelligent Systems, University of Passau, Germany
e.coutinho@imperial.ac.uk

Abstract

In this paper, we describe a new database with audio recordings of non-native (L2) speakers of English, and the perceptual evaluation experiment conducted with native English speakers for assessing the prosody of each recording. These annotations are then used to compute the gold standard using different methods, and a series of regression experiments is conducted to evaluate their impact on the performance of a regression model predicting the degree of naturalness of L2 speech. Further, we compare the relevance of different feature groups modelling prosody in general (without speech tempo), speech rate and pauses modelling speech tempo (fluency), voice quality, and a variety of spectral features. We also discuss the impact of various fusion strategies on performance. Overall, our results demonstrate that the prosody of non-native speakers of English as L2 can be reliably assessed using supra-segmental audio features; prosodic features seem to be the most important ones.

Keywords: Index Terms: non-native speech, prosody, feature evaluation

1. Introduction

Non-native (L2) speakers of English diverge from native English speakers in terms of linguistic (e.g, morphology, syntax, lexicon) and phonetic aspects. Phonetic aspects comprise segmental and supra-segmental (prosodic) phenomena. There are two main tasks within automatic speech processing where an assessment of non-native traits can help. First, the performance of Automatic Speech Recognition (ASR) degrades drastically while processing non-native speech. A central aspect here is the mismatch between training data from native speech and test data from non-native speech (Van Compernolle, 2001), preventing the acoustic models to capture the variability and 'innovations' introduced by L2 speakers. By assessing the degree of nativeness in general, and by specifying the speaker's L1, it is possible to build speaker- and/or speaker-group specific acoustic models. Second, such an assessment can be employed within Computer-Aided Pronunciation Training (CAPT) for the automatic screening and assessment of learners.

In this paper, we describe a new database with audio recordings of L2 English speakers, and the perceptual evaluation experiment conducted with native English speakers to obtain judgements of the naturalness of the pronunciation of each recording with respect to supra-segmental characteristics. These annotations are then used to compute the gold standard using different methods, and a series of regression experiments is conducted to evaluate their impact on the performance of a regression model predicting the degree of naturalness of L2 speech. We use the publicly available general-purpose feature set openSMILE (Eyben et al., 2013) to compute a large set of acoustic features. Additionally, to add some more specialised (but still text-independent features) we also compute duration, rhythm, and prosodic features derived from a segmentation of the recordings into vocalic and consonantal intervals, and from pseudo-syllables derived from that segmentation. Those vocalic and consonantal intervals were inferred from the output of a phoneme recognizer (PR) (Schwarz et al., 2006).

2. Recording and annotation methods

2.1. Collection of voice recordings

We obtained a set of voice recordings from 54 non-native English speakers with varied degree of proficiency (see Table 1 for details). Each speaker was asked to read aloud a set of 11 sentences from two short stories written in the English language widely used in phonetics and speech pathology research. The data recorded consisted of 5 sentences taken from the fable "The North Wind and the Sun", and 6 sentences extracted from "The Rainbow". All recording sessions took place in a room at the Technische Universität München (Germany) using the same hardware and software. The full database comprises 594 recordings (11 sentences * 54 speakers), totalling to 1.4 hours of speech.

2.2. Annotation procedure

The full set of recordings was delivered for annotation to a group of 27 native English speakers. The group of annotators were instructed to rate the naturalness of supra-segmental characteristics by answering the question "*The English language has a characteristic prosody (sentence melody and rhythm, i. e. timing of syllables). This sentence's prosody sounds ...*," using a 5-point Likert scale (1 - normal; 2 - acceptable; 3 - slightly unusual; 4 - unusual;

Mother tongue (L1)
Arabic (4), Bulgarian (1), Chinese (13), French (1), German (22), Italian (2), Indian (2), Japanese (1), Lithuanian (1), Persian (2), Polish (1), Portuguese (1), Romanian (1), Russian (1), Turkish (1)
Age
M = 31 y.o.; SD = 9 y.o; Range = 20-57 y.o.
Gender
28 female / 26 male

Table 1: *Non-native (L2) English speakers' characteristics. M - mean; SD - standard deviation*

5 - very unusual). [1] Inter-rater agreement (per session and rater gender) was computed using the Intraclass correlation (ICC; (Shrout and Fleiss, 1979)). In particular, we computed $ICC(2,k)$ as described by Shrout and Fleiss (1979), which estimates the absolute agreement between subjects for k ratings. Results are shown in the last column of Table 2.

Session	#recordings	#raters	$ICC(2,k)$
1	132	23 (All)	0.96
		15 (Female)	0.94
		8 (Male)	0.89
2	132	16 (All)	0.92
		10 (Female)	0.86
		6 (Male)	0.84
3	165	17 (All)	0.93
		11 (Female)	0.88
		6 (Male)	0.87
4	165	18 (All)	0.95
		11 (Female)	0.89
		7 (Male)	0.91
mean	-	19 (All)	0.94
		12 (Female)	0.89
		7 (Male)	0.88

Table 2: *Rating sessions details: number of raters (#raters), number of recordings (#recordings) rated in each session, and inter-rater agreement per session ($ICC(2,k)$ - Intraclass correlation).*

2.3. Gold standard

We tested three different methods to infer the gold standard from individual annotations. First, we combined the individual ratings for each instance by calculating the median across raters for each instance. Second, we computed the arithmetic mean across all raters for each instance. Third, in order to consider individual (unknown) differences amongst annotators (e.g., level of attention or individual biases when providing the ratings), we attempted to improve the quality of the gold standard calculation by using the evaluator weighted estimator (EWE) method as described in (Grimm and Kroschel, 2005). The EWE average of the individual ratings considers that each evaluator is subject to an individual amount of disturbance during the evaluation, by introducing evaluator-dependent weights that correspond to the correlation between the listener's responses and the average ratings of all evaluators. The histogram of frequencies for each gold standard is shown in Fig. 1.

[1] This is a sort of 'unified' rating of both melody and rhythm at the same time – in (Hönig et al., 2012a), it was shown that independent melody and rhythm assessments correlate highly with each other. Melody is a rather straightforward concept which has been dealt with in many different intonation models. Rhythm is a rather vague concept and could not defined strictly so far; yet, we have seen that it can be conveyed and used within such rating experiments.

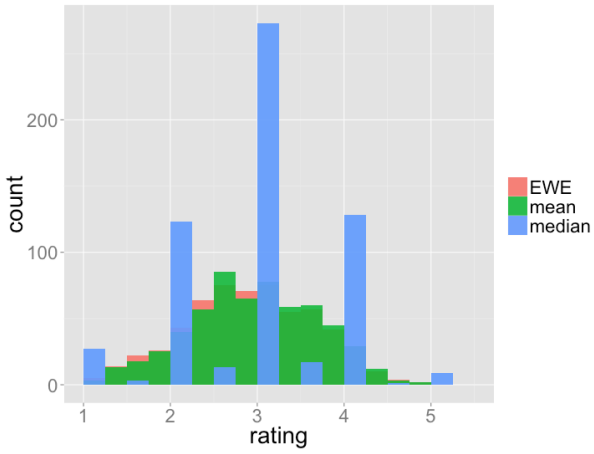

Figure 1: *Histogram of frequencies for each gold standard.*

3. Feature sets

We used five different features sets for acoustic modelling of this database as described in the following paragraphs.

OpenSMILE Features (*oS*): openSMILE (version 2.0; (Eyben et al., 2013)) is a framework for extracting general-purpose acoustic and prosodic features, and has been applied successfully to a garden-variety of paralinguistic tasks. In this paper we employ the configuration file of the INTERSPEECH 2013 Computational Paralinguistics ChallengE (*ComParE*) (Schuller et al., 2013), which includes 6373 static features of functionals of low-level descriptor (LLD) contours. In order to evaluate the nature of acoustic features that permit to better predict the L2 speakers' prosodic competence, we subdivided the *ComParE* set into three partitions: prosodic (*oS_PROS*), voice quality (*oS_VQ*), and spectral and cepstral features (*oS_SPEC*). The LLDs in each subset are described in (Weninger et al., 2013).

Prosodic Features (*PROS*): The set of features extracted (a total of 64) are based on duration, energy, pitch, and pauses, and are designed to locally describe arbitrary units of speech such as words or syllables (Batliner et al., 2000). Here, we used the pseudo-syllables derived from the phoneme recognizer. The sequence of these local features was then converted into a fixed-length vector using 12 functionals (e.g., mean, standard deviation).

Rhythm Features (*RHYT*): Given the importance of temporal features to the evaluation of L2 speakers' proficiency (cf. Section 1.), we considered a set of specific speech-rhythm related features. Based on the segmentation of

the PR (see above) into vocalic and consonantal intervals, we computed Grabe's raw pairwise variability index (rPVI; (Grabe and Low, 2002)) on consonants and vowels plus its (rate-of-speech-) normalised version (nPVI). Additionally, we computed variants of Grabe's measures using squared instead of absolute differences (root average squared local change). This led to 8 features reflecting local variability in durations. Additionally, another set of five features was added to reflect global variability and proportions. These were calculated using Ramus' V (percentage of vocalic intervals) and vocalic and consonantal Deltas (global standard deviations of durations) (Ramus, 2002), plus Dellwo's variation coefficient Varco (rate-of-speech-normalised standard deviation of durations) (Dellwo, 2010). Rhythmic features have also been used in previous speech scoring research (Chen and Zechner, 2011).

Duration Features (*DUR*): In order to include temporal features specifically modelling speech tempo, we created another set of features comprising *articulation rate* (the number of syllables divided by the total length of the speech segments) and the *percentage of pauses* (excluding initial and final silences in each recording). These features were computed using the segmentation of pauses, vowels, consonants, and speaker noise derived from the PR.

4. Experiments and Results

In this section we describe the results of the acoustic modelling of our database using the features described in Section 3.. Additionally, we also present the results from various combinations of the individual feature sets using early and late fusion approaches. For early fusion, we just use the superset of the individual feature sets. For late fusion, a regressor is trained for each individual feature set, and the outputs are combined with another regressor. In all our experiments we employ Support Vector Regression (SVR) (Smola and Schlkopf, 2004) with linear kernel and Sequential Minimal Optimization learning (SMO) (Platt, 1999). In all experiments (implemented with *sklearn* (Pedregosa et al., 2011) and *LibSVM* (Chang and Lin, 2011)), the performance of the models was estimated using a 10-fold speaker-independent cross-validation scheme (SICV), i. e., the train and test set of each fold was disjunct w. r. t. speakers. The SVR complexity parameter C was optimised in each fold using an inner 10-fold SICV on the training set of that fold. In case of late fusion, the predictions obtained with the help of that inner SICV were re-used for training the final regressor. The SVR performance was estimated using Spearman's rank correlation coefficient between target and predicted values on all test instances (ρ). The details and results of our tests are shown in Tables 3 (individual features sets) and 4 (fusion). The best results using individual features sets were obtained with the *PROS* ($\rho = .600$), *oS_PROS* ($\rho = .542$) and *DUR* ($\rho = .526$) feature sets. This confirms our initial expectations (see Section 1.) regarding the relevance of prosodic features (including speech tempo) to the perception of naturalness of L2 speech. In relation to the fusion tests (see Table 4), we evaluated various combinations of the individual sets. Chiefly, we found that (1) late fusion was globally better than early fusion (av. $\rho(early) = .562$; av. $\rho(late) = .586$); (2) the best performance was obtained with the late fusion of all feature sets except *oS_SPEC*; (3) the top performance is only slightly higher (.017 increase in ρ) than the performance of the best individual set *PROS* ($\rho = .600$).

Feature set	# features	ρ
oS_PROS	483	0.542
oS_SPEC	5500	0.514
oS_VQ	390	0.513
DUR	2	0.526
PROS	3072	**0.600**
RHYT	13	0.411

Table 3: *Regression experiments results for the EWE gold standard using individual features sets.*

Feature set	ρ	
	early fusion	*late fusion*
oS_PROS+oS_SPEC	0.508	0.544
oS_PROS+oS_VQ	0.551	0.552
oS_PROS+DUR	0.547	0.571
oS_PROS+PROS	0.600	0.602
oS_PROS+RHYT	0.544	0.573
oS_SPEC+oS_VQ	0.519	0.543
oS_SPEC+DUR	0.515	0.566
oS_SPEC+PROS	0.576	0.603
oS_SPEC+RHYT	0.513	0.546
oS_PROS+DUR	0.526	0.569
oS_VQ+PROS	0.607	0.606
oS_VQ+RHYT	0.607	0.567
DUR+PROS	0.600	0.604
DUR+RHYT	0.554	0.565
PROS+RHYT	0.599	0.604
all	0.572	0.616
all-oS_PROS	0.575	0.614
all-oS_SPEC	**0.608**	**0.617**
all-oS_VQ	0.573	0.612
all-DUR	0.572	0.608
all-PROS	0.522	0.599
all-RHYT	0.572	0.604

Table 4: *Regression experiments results for the EWE gold standard using different combinations of the individual feature sets (early and late fusion approaches).*

5. Discussion and Conclusions

In this paper, we presented a new database with audio recordings of non-native (L2) speakers of English, and the perceptual evaluation experiment conducted with native English speakers for assessing the prosody of each recording. These annotations were used to compute the gold standard using different methods, and a series of regression experiments to evaluate their impact on the performance of a regression model predicting the degree of naturalness of L2 speech. To this end, we compared the relevance of different feature groups modelling prosody in general (without speech tempo), speech rate and pauses modelling speech tempo (fluency), speech rhythm, voice quality, and a variety

of spectral features. Overall, using a speaker-independent schema, we have demonstrated that it is feasible to predict the level of naturalness of L2 speech using acoustic features alone. More specifically, we corroborated the findings from (Hönig et al., 2010; Hönig et al., 2012a; Hönig et al., 2012b) that specialised prosodic features (including speech tempo) seem to be the relevant ones to assess the naturalness of English produced by non-native speakers. Naturally enough, the two features modelling speech tempo (*DUR*) alone already yield a very good performance because speech tempo models fluency, and fluency is highly correlated with the degree of nativeness.

The question remains which other features and methods could be employed in addition, and which might be the 'empirical' upper baseline for the performance. For the upper baseline, we computed the averaged gold standard per speaker, under the simplifying assumption that speaker performance is constant within the same recording session. The correlation with the EWE gold standard was ρ=0.823. When we employ the default cross-validation procedures of toolboxes such as WEKA or sklearn, we obtain (with 10 folds and late fusion of all six feature groups) ρ=0.715. However, these procedures are not speaker-independent. These two figures indicate the range of improvement that we can end up with when employing (1) item specific information obtained via speech recognition and (2) speaker modelling.

6. Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378 (ARIA-VALUSPA), the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant 01IS07014B (C-AuDiT), and the German Ministry of Economics (BMWi) under grant KF2027104ED0 (AUWL), and the European Union's 7th Framework Program under grant agreement No. 338164 (Starting Grant iHEARu).

7. Bibliographical References

Batliner, A., Buckow, J., Niemann, H., Nöth, E., and Warnke, V. (2000). The prosody module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, pages 106–121. Springer, Berlin.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, L. and Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. In *Proc. INTERSPEECH 2011*, pages 1861–1864.

Dellwo, V. (2010). Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. *Universität Bonn, Bonn University. Bonn, Germany*.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*, pages 835–838, Barcelona, Spain, October. ACM, ACM.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7(515-546).

Grimm, M. and Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *Proc. ASRU*, pages 381–385. IEEE.

Hönig, F., Batliner, A., Weilhammer, K., and Nöth, E. (2010). Automatic assessment of non-native prosody for English as l2. In *Proc. Speech Prosody*, Chicago. no pagination.

Hönig, F., Batliner, A., and Nöth, E. (2012a). Automatic assessment of non-native prosody – annotation, modelling and evaluation. In *Proceedings of IS-ADEPT, International Symposium on Automatic Detection of Errors in Pronunciation Training, June 6-8, Stockholm, Sweden*.

Hönig, F., Bocklet, T., Riedhammer, K., Batliner, A., and Nöth, E. (2012b). The automatic assessment of non-native prosody: Combining classical prosodic analysis with acoustic modelling. In *Proc. INTERSPEECH 2012*, pages 823–826, Portland (Oregon),USA, September.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods – support vector learning*, pages 185–208. MIT Press.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: perspectives. In *Proc. of Speech Prosody*, pages 115–120, Aix-en-Provence, France. Laboratoire Parole et Langage.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. INTERSPEECH 2013*, pages 148–152, Lyon, France, August. ISCA, ISCA.

Schwarz, P., Matejka, P., and Cernocky, J. (2006). Hierarchical structures of neural networks for phoneme recognition. In *Proceedings of ICASSP*, pages 325–328, Toulouse.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Smola, A. and Schlkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Van Compernolle, D. (2001). Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication*, 35(1):71–79.

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., and Scherer, K. R. (2013). On the Acoustics of Emo-

tion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Psychology*, 4:1–12.