Language Proficiency Assessment of English L2 Speakers Based on Joint Analysis of Prosody and Native Language

Yue Zhang
Department of Computing
Imperial College London
London, U.K.
yue.zhang1@imperial.ac.uk

Felix Weninger
Nuance Communications
Ulm, Germany
felix@weninger.de

Anton Batliner
Chair of Complex and
Intelligent Systems
University of Passau
Passau, Germany
batliner@cs.fau.de

Florian Hönig
Pattern Recognition Lab
FAU Erlangen-Nuremberg
Erlangen, Germany
hoenig@informatik.uni-erlangen.de

Björn Schuller
Department of Computing
Imperial College London
London, U.K.
bjoern.schuller@imperial.ac.uk

ABSTRACT

In this work, we present an in-depth analysis of the interdependency between the non-native prosody and the native language (L1) of English L2 speakers, as separately investigated in the Degree of Nativeness Task and the Native Language Task of the INTERSPEECH 2015 and 2016 Computational Paralinguistics ChallengE (ComParE). To this end, we propose a multi-task learning scheme based on auxiliary attributes for jointly learning the tasks of L1 classification and prosody score regression. The effectiveness of this approach is demonstrated in extensive experimental runs, comparing various standardised feature sets of prosodic, cepstral, spectral, and voice quality descriptors, as well as automatic feature selection. In the result, we show that the prediction of both prosody score and L1 can be improved by considering both tasks in a holistic way. In particular, we achieve an 11 % relative gain in regression performance (Spearman's correlation coefficient) on prosody scores, when comparing the best multi- and single-task learning results.

CCS Concepts

•**Computing methodologies → Multi-task learning;**

Keywords

Non-Native Prosody; L1 Identification; Feature Evaluation

1. INTRODUCTION

As spoken language applications become more frequent in global business and commerce, it is envisioned that systems performing both language proficiency assessment and native language identification will be increasingly in demand. For example, context-aware spoken dialogue systems can exploit accent-specific acoustic models, adapt the tempo of speech synthesis to the language proficiency of individual speakers, or even switch to a user's native language in case of difficulties with the interaction in the default language. Realising these capabilities in automatic systems conceivably leads to more natural and human-like interaction, as humans typically adapt to the language proficiency of their counterparts, as well as better recognition accuracy and user customisation. Another relevant application field is Computer-Assisted Pronunciation Training (CAPT) for providing corrective feedbacks to language learners [28, 16].

Non-native speakers of English diverge from native English speakers in terms of linguistic (e. g., morphology, syntax, lexicon) and phonetic aspects, comprising segmental and supra-segmental (prosodic) traits. Located on word level and above, prosodic speech phenomena encompass word accent position, syntactic-prosodic boundaries, and rhythm, hence determine the language proficiency in a second language (L2) and by that, mutual understanding [13].

In the *Degree of Nativeness task* of the Computational Paralinguistic ChallengE (ComParE) 2015 [23], prosody with respect to sentence melody and rhythm was considered to assess the pronunciation quality of English L2 speakers on a rating scale. Further studies targeting in particular this task include, e. g., [13, 25, 26]. In contrast, in the ComParE 2016 *Native Language task* [24], the native language (L1) of non-native English L2 speakers from eleven L1 backgrounds has to be automatically identified. Relatedly, a few studies investigated non-native accent identification using prosodic parameters [17, 15], and supra-segmental native traits when trying to model language-specific rhythm [20]. However, in the literature there has been little work on native language identification despite its similarity to the task of language identification (in which a system distinguishes between different languages), and dialect or accent identification (in which a system recognises regional native speaker dialects of a single spoken language, such as British vs American

English ([2, 3, 14]).

Most importantly, the interdependency between non-native prosody and L1 background has not been considered, let alone exploited for automatic speech analysis, although there exists an intuitive and proven coherence between them [6]. Overall, L1 background was identified to have a high impact on the degree of L2 foreign accent, besides other factors such as age of L2 learning, length of residence in an L2-speaking country, gender, formal instruction, language learning aptitude and motivation, as well as amount of native language (L1) use [18]. Linguistically and phonetically speaking, the global language system can be divided into different language groups. For example, a difference in rhythm in terms of isochrony can be observed between syllable timed languages such as French, and stress-timed languages such as English [1]. More recent studies [9, 20] indicate that syllables that are weak in stress-timed languages are pronounced stronger in syllable-timed languages, resulting in L1-specific idiosyncrasies and prosodic traits, respectively. Therefore, depending on the resemblance of two languages, it seems reasonable to assume that English L2 speakers from different L1 backgrounds produce more or less 'natural' pronunciation. That is to say, the accuracy with which nonnative speakers pronounce an L2 is, at least to some extent, dependent on their L1 [18].

Motivated by this hypothesis, we conducted extensive studies on the automatic analysis of non-native speech, showing that the recognition performance for each task can be significantly improved by jointly learning both speaker traits.

2. METHODOLOGY

2.1 Multi-Task Learning

Let us first introduce some required notation: $\mathbf{x}_i^{(l)} \in \mathcal{X}^{(l)}$ denotes the i-th feature vector for classification task l, while $y_i^{(l)} \in \mathcal{Y}^{(l)}$ denotes its gold standard label for task l, where $\mathcal{X}$ is the acoustic feature space and $\mathcal{Y}^{(l)}$ is the label space for task l. $[\cdot;\cdot]$ denotes the concatenation of features. To exploit the interdependency between non-native prosody and L1 of English L2 speakers, we use auxiliary features in a classifier chain, similar to the work [21]. The idea is to append 'L1' as attribute to the feature vector associated with the target label 'prosody' in the Degree of Nativeness task (DN), and vice versa for the Native Language task (N). In case that the auxiliary label is missing in a specific database, we have to rely on a predicted label obtained by semi-supervised learning, exploiting a classifier $\tilde{y}^{(m)} : \mathcal{X} \rightarrow \mathcal{Y}^{(m)}$ trained on the original training set of task l. In the scope of this paper, $l, m \in \{\mathrm{DN}, \mathrm{N}\}$. Accordingly, the multi-task training data can be formularised as pairs of feature vectors and labels $[x_i^{(l)}; \tilde{y}_i^{(m)}], y_i^{(l)}$. Otherwise, if a database already has a gold standard $y_i^{(m)}$ for the auxiliary label, we can use this as attribute. Here, it is noted that this approach is a form of Cross-Task-Labelling (CTL) as proposed in our previous work [31, 30], which can be understood as a generalisation of semi-supervised learning to L-dimensional labels, where each dimension corresponds to a classification task. As the multi-task learning scheme is based on combining acoustic features with labels in the feature space, it is crucial to evaluate the performance with a variety of acoustic feature sets to assess the complementarity.

2.2 Feature Sets

For the audio and acoustic analysis, we compare the relevance of six feature groups. A full description of the feature groups can be found in [27, 22].

ComParE: The *ComParE* set of supra-segmental acoustic features is a well-evolved set for automatic recognition of paralinguistic speech phenomena, as used for the baseline of the INTERSPEECH ComParE series. It contains 6 373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours. The configuration file is included in the 2.1 public release of openSMILE [8, 7]. Important subgroups of the ComParE feature set comprise prosodic (*PROS*), Mel Frequency Cepstral Coefficients (*MFCC*), spectral (*SPEC*), and voice quality (*VQ*) features.

Prosodic Features (*PROS*): Given the importance of sentence melody and rhythm to the assessment of L2 speakers' proficiency (cf. Section 1), we extracted a set of prosodic features based on loudness, energy, and F_0 (SHS & Viterbi smoothing) to locally describe arbitrary units of speech such as words or syllables.

Mel Frequency Cepstral Coefficients (*MFCC*): MFCC features are among the most popular speech features for automatic speech recognition, music information retrieval, and a wide variety of paralinguistic tasks. The mel scale takes human hearing perception into account, where lower frequencies are resolved better by human hearing than higher ones [5].

Spectral Features (*SPEC*): Spectral statistical LLDs, such as spectral variance and spectral flux, are often used in multimedia analysis, and are part of the descriptor set proposed in the MPEG-7 multimedia content description standard. For this reason, they are very relevant for music and sound analysis.

Voice Quality (*VQ*): Voice quality LLDs are used to discriminate harmonic and noise-like sounds. Like the harmonics-to-noise ratio (HNR), they describe the quality of the excitation signal and thus the quality of the voice. For instance, jitter and shimmer are micro-prosodic variations of the length and amplitudes of the fundamental frequency for harmonic sounds.

Correlation-based Feature Selection (*CFS*): The brute-force combination of LLDs and functionals potentially yields features that are irrelevant for the machine learning task considered, or not meaningful in general. Automatic feature selection is a data-driven way to deal with this problem, such as by selecting features that exhibit high correlation with the target label(s). However, since feature brute-forcing yields many features of similar nature and hence similar correlation with the target label(s), it is required to combine feature selection with feature decorrelation to arrive at a small, yet efficient feature space. To this end, correlation-based feature selection (CFS) can be employed [29]. There, the merit M of a feature subset S with k features is given by

$$M(S) = \frac{k \ \mathrm{CC}_{\mathrm{cf}}}{\sqrt{k + k(k-1)\mathrm{CC}_{\mathrm{ff}}}}, \qquad (1)$$

where $\mathrm{CC}_{\mathrm{cf}}$ denotes the mean correlation coefficient (CC) of features in S with the class label, and $\mathrm{CC}_{\mathrm{ff}}$ is the average CC of features in S with each other. It is easy to see that a candidate subset that maximises M will provide an optimal trade-off between high predictive power regarding the class label (numerator) and low redundancy among the features (denominator).

Table 1: Databases of Non-Native Spoken English*: Number of instances per class in the train/devel/test split used for the Challenges.*

#	*ComParE 2015*			*ComParE 2016*			
	Train	Test	Σ	Train	Devel	Test	Σ
ARA	62	-	62	300	86	80	466
CHI	203	-	203	300	84	74	458
FRE	257	166	423	300	80	78	458
GER	2 208	451	2 659	300	85	75	460
HIN	-	37	37	300	83	82	465
HUN	216	-	216	-	-	-	-
ITA	290	176	466	300	94	68	462
JAP	406	-	406	300	85	75	460
KOR	-	-	-	300	90	80	470
POR	248	-	248	-	-	-	-
SPA	-	169	169	300	100	77	477
TEL	-	-	-	300	83	88	471
TUR	-	-	-	300	95	90	485
Σ	3 890	999	4 889	3 300	965	867	5 132

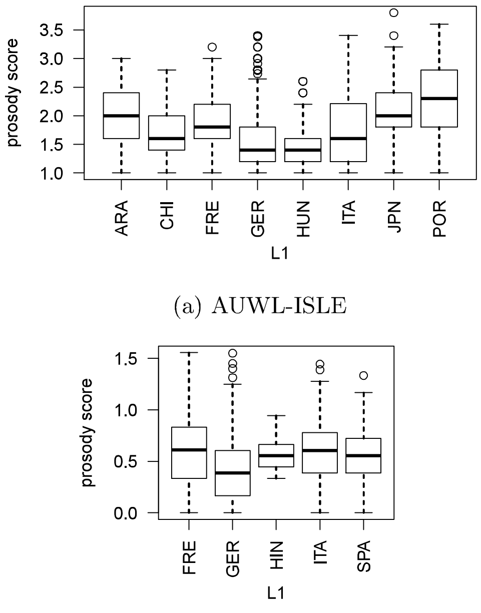
3. CORPORA

3.1 ComParE 2016

The Educational Testing Service (ETS) Corpus of Non-Native Spoken English comprises more than 64 hours of speech from 5 132 non-native speakers of English, with eleven different L1 backgrounds (Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR); cf. Table 1). Each speech recording is 45 seconds long and was obtained in the context of the TOEFL iBT assessment, which is designed to measure a non-native speaker's ability to use and understand English at the university level. The dataset was divided into speaker-independent partitions: 3 300 instances (64 %, approximately 41.3 hours) were selected as training data, 965 instances (19 %, approximately 12.1 hours) for the validation set, and 867 responses (17 %, approximately 10.8 hours) were used as test data.Unlike the ComParE 2015 databases, a prosody score is not provided and hence, CTL is used to obtain the missing labels by machine labelling.

3.2 ComParE 2015

Following the protocol of the ComParE 2015, the AUWL and ISLE databases are used for training while the C-AuDiT database is used for testing. The AUWL corpus [11] comprises 31 speakers (13 f, 18 m; 36.5 ± 15.3 years; native languages: 16 GER, 4 ITA, 3 CHI, 3 JAP, 2 ARA, 1 Portuguese (POR), 1 Hungarian (HUN); cf. Table 1), 5.5 hours, 3 732 speech files (423 distinct sentences/phrases). Each speech file was annotated by five phoneticians with respect to its prosody (sentence melody and rhythm) on a five-point scale ranging from (1) for normal to (5) for very unusual. With the (simplifying) assumption of an interval scale, we took the arithmetic average of the five labellers to obtain inter-subjective prosody scores [13], with an average of 1.7 and a standard deviation of 0.5 (range 1.0–3.8). From the ISLE corpus, we used material comprising 36 speakers (11 f, 25 m; native languages: 20 GER, 16 ITA), 0.3 hours, 158 speech files (5 distinct sentences); prosody scores were collected in a similar manner (2.1 ± 0.5, range 1.3–3.4). The C-AuDiT database [12] contains read non-native English short stories broken down into single sentences. The material comprises 58 speakers (31 f, 27 m; native languages: 26 GER, 10 FRE, 10 ITA, 10 SPA, 2 HIN), 2.7 hours, and 999 speech files (19 distinct sentences). Prosodic scores were collected similarly to AUWL, except for using a 3-point scale from 0 for good to 2 for bad (0.5 ± 0.3, range 0.0–1.6). Figure 1 shows the distribution of prosody scores for each L1 in the AUWL, ISLE, C-AuDiT (AIC) database family.



(a) AUWL-ISLE

(b) C-AUDIT

Figure 1: Distribution of prosody scores in the AUWL-ISLE and C-AUDIT databases.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Setup

In our experiments, we applied Support Vector Machines (SVM) with linear kernels for the classification task, and Support Vector Regression (SVR; also with linear kernels) with epsilon-insensitive loss for regression. In all tasks, the Sequential Minimal Optimisation (SMO; [19]) algorithm was used for training. For transparency and reproducibility, we used open-source implementations (WEKA 3, revision 3.7.13; [10]. The ϵ parameter was fixed at 1.0, while the complexity parameter C was optimised by training set cross-validation. Due to different text materials and recording conditions, the latter did not always result in optimal values for test. Features were scaled to zero mean and unit standard deviation, using the scales and offsets from the training set. In training set cross-validation, these were calculated on the training set of each fold. Considering the cross-corpus nature of the Degree of Nativeness task, we used a 4-fold double nested loop over speakers and texts for cross-validation as defined in the ComParE 2015 [23].

Moreover, we separately evaluated the performance on the test set containing only the intersection set of the L1 backgrounds found in the AUWL, ISLE, and C-AuDiT databases

Table 2: ***Degree of Nativeness (DN) and Native Language (N); Performance on the test set in terms of ρ (Spearman's rank correlation) and unweighted average recall (UAR), by single-task learning (ST), and by multi-task learning (MT).***

Feature set	#	DN (ρ)		N (UAR[%])	
		ST	MT	ST	MT
ComParE	6 373	.427	.447	47.5	**47.8**
SPEC	2 600	.326	.335	40.8	40.0
MFCC	1 400	.412	.444	39.4	39.8
PROS	483	.366	.474	33.5	34.8
VQ	390	.340	.382	26.8	25.5
CFS	98	.409	**.476**	41.2	42.1

Table 3: ***Degree of Nativeness (DN) performance by ρ for various feature sets (#: number of features) with ST, MT, as well as dummy regression corresponding to training mean prediction, on a subset of the C-AUDIT database comprising speakers whose L1 background is also found in the AUWL-ISLE set.***

Feature set	#	ST	MT
SVR on acoustic features			
ComParE	6 373	.435	.459
SPEC	2 600	.338	.348
MFCC	1 400	.416	.458
PROS	483	.402	**.511**
VQ	390	.364	.409
CFS	98	.431	.495
Training mean prediction			
mean pros.	1	0.273	

(FRE, GER, and ITA), for a total of 793 instances. This serves to verify the hypothesis that the performance depends on whether the L1 background of a test speaker has already been seen in training or not. For the regression task, we considered Spearman's rank correlation coefficient ρ, given the ordinal-scaled annotations as outlined in Section 3, as in the ComParE 2015. The evaluation measure for the classification task is unweighted average recall (UAR), as per the ComParE 2016 protocol.

4.2 Discussion of Results

In this section, we describe the results of the acoustic modelling using the feature sets described in Section 2.2, where the baseline is defined as the accuracy achieved by ST learning when using the ComParE set. From the results in Table 2, it can be clearly seen that on the Degree of Nativeness task, MT consistently outperforms ST for each feature set considered. In particular, the best MT result yields an 11 % relative gain in ρ over the best ST result. The performance gain by MT over ST is largest for the PROS feature set (30 % relative).

As regards the performance of different feature sets, we observe feature selection by CFS to be particularly effective for prosody score regression. The relevance of the auxiliary attributes is highlighted by the fact that they are selected by the CFS method out of 6 373 ComParE features, remaining as one of the 136 features for L1 classification and 98 for prosody score regression, respectively. Using only 98 features, we are able to improve over the large ComParE feature set by 11 % relative. Further, we corroborate the findings from [4] that prosodic features seem to be appropriate to assess the naturalness of English produced by non-native speakers.

In contrast, only a slight improvement tendency by MT can be observed for the Native Language task. Investigating this further, we found that the predicted prosody scores on the ETS corpus were all in a tight range. This can be explained by a comparable English proficiency of all the TOEFL examinees considered in the study, but it is also conceivable that it might be due to low performance of the prosody regression on the ETS corpus. Interestingly, we also found that the various feature groups ranked differently as to their worth for prosody regression vs L1 classification.

In Table 3, we show the results for the Degree of Nativeness task obtained on the reduced test set of L1 backgrounds that match the training set L1 backgrounds. As expected, the ST performance is higher on this test set than on the full test set, due to a larger match of training and test data. However, the gain by MT is similar on both test sets, and in the result, we are able to achieve up to .511 rank correlation by prosodic features and MT learning. For prosodic features, the gain by MT over ST is significant according to a z-test ($p < .001$). In Table 3, we also show the performance of a 'dummy' regressor that, given the L1 of a test speaker, outputs the mean prosody score among the training speakers with the same L1. This achieves a ρ of .273, which is significantly lower ($p < .005$ according to a z-test) than any of the performances with acoustic features achieved on the same test set. We can thus conclude that the notable regression performances by acoustic features cannot be simply explained by implicit L1 classification. It remains to investigate if the large gain by including L1 is due to the regressor simply learning the mean score of each L1 group, which could be conjectured due to the unequal distributions per L1 group (cf. Figure 1). To this end, we conducted an additional evaluation of Spearman's ρ within each L1 group in the test set, using the PROS features. We found an improved ρ by MT within each of the French, German, and Italian L1 groups, although the difference was smaller in magnitude (up to .044) than the total gain by MT on the whole test set (.109). This confirms that the L1 feature helps the regression beyond a simple mean adjustment per L1 group.

5. CONCLUSIONS

In a large-scale study on the interdependency of automatic prosody assessment and L1 background identification of non-native English speakers, we were able to confirm the hypothesis that knowledge of the L1 background of non-native English speakers helps significantly in assessing their language proficiency and vice versa. In future work, we will extend our multi-task learning scheme to incorporate automatic prediction of L1 into prosody score regression, as well as to use other learning models such as shared-hidden-layer deep neural networks.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Framework Programme for Research and Innovation HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA). We further thank ELRA for providing the ISLE Speech Corpus (ref. ELRA-S0083).

7. REFERENCES

[1] D. Abercrombie. *Elements of general phonetics*, volume 203. Edinburgh University Press Edinburgh, Chicago, IL, 1967.

[2] L. M. Arslan and J. H. Hansen. Language accent classification in american english. *Speech Communication*, 18(4):353–367, 1996.

[3] F. Biadsy. *Automatic dialect and accent recognition and its application to speech recognition.* PhD thesis, Columbia University, 2011.

[4] E. Coutinho, F. Hönig, Y. Zhang, S. Hantke, A. Batliner, E. Nöth, and B. Schuller. Assessing the prosody of non-native speakers of english: Measures and feature sets. In *Proc. of LREC*, pages 1328–1332, Portoroz, Slovenia, 2016. ELRA.

[5] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[6] A. De Swaan. *Words of the world: The global language system.* John Wiley & Sons, 2013.

[7] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. of ACM Multimedia*, pages 835–838, Barcelona, Spain, 2013. ACM.

[8] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of ACM Multimedia*, pages 1459–1462, Florence, Italy, 2010. ACM.

[9] E. Grabe and E. L. Low. Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7, 2002.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, 2009.

[11] F. Hönig, A. Batliner, and E. Nöth. Automatic assessment of non-native prosody – annotation, modelling and evaluation. In *Proc. of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, pages 21–30, Stockholm, 2012.

[12] 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 *Proc. of SLATE*, Wroxall Abbey, 2009.

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

[14] L. Kat and P. Fung. Fast accent identification and accented speech recognition. In *Proc. of ICASSP*, volume 1, pages 221–224, Phoenix, Arizona, 1999. IEEE.

[15] J. Lopes, I. Trancoso, and A. Abad. A nativeness classifier for ted talks. In *Proc. of ICASSP*, pages 5672–5675, Prague, Czech Republic, 2011. IEEE.

[16] A. Neri, C. Cucchiarini, and H. Strik. ASR-based corrective feedback on pronunciation: does it really work? In *Proc. of Interspeech*, pages 1818–1821, Pittsburgh, PA, 2006. ISCA.

[17] M. Piat, D. Fohr, and I. Illina. Foreign accent identification based on prosodic parameters. In *Proc. of Interspeech*, pages 759–762, Brisbane, Australia, 2008. ISCA.

[18] T. Piske, I. R. MacKay, and J. E. Flege. Factors affecting degree of foreign accent in an l2: A review. *Journal of phonetics*, 29(2):191–215, 2001.

[19] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999.

[20] F. Ramus. Acoustic correlates of linguistic rhythm: Perspectives. In *Proc. of Speech Prosody*, pages 115–120, Aix-en-Provence, 2002.

[21] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

[22] B. Schuller and A. Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing.* Wiley, 2013.

[23] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Höth, Y. Zhang, and F. Weninger. The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition. In *Proc. of Interspeech*, pages 478–482, Dresden, Germany, 2015. ISCA.

[24] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception & Sincerity. In *Proc. Interspeech*, San Francsico, CA, 2016. ISCA. 2001–2005.

[25] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and M. K. Sönmez. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *Proc. of Interspeech*, Beijing, 2000. 187–190.

[26] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom. Testing suprasegmental english through parroting. In *Proc. of Speech Prosody*, Chicago, IL, 2010.

[27] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer. On the acoustics of emotion in audio: What speech, music and sound have in common. *Frontiers in Emotion Science*, 4(292):1–12, 2013.

[28] S. M. Witt. *Use of speech recognition in computer-assisted language learning.* Ph.D. dissertation, University of Cambridge, 1999.

[29] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[30] Y. Zhang, F. Weninger, Z. Ren, and B. Schuller. Sincerity and deception in speech: Two sides of the same coin? A transfer- and multi-task learning perspective. In *Proc. of Interspeech*, pages 2041–2045, San Francisco, CA, 2016. ISCA.

[31] Y. Zhang, Y. Zhou, J. Shen, and B. Schuller. Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis. In *Proc. of ICASSP*, pages 6090–6094, Shanghai, P. R. China, 2016. IEEE.