Intonation-Based Classification of Language Proficiency Using FDA

Oliver Jokisch[1], *Tristan Langenberg*[1], *Gábor Pintér*[2]

[1]Institute of Communications Engineering, Leipzig University of Telecommunication, Germany
[2]School of Languages and Communication, Kobe University, Japan

jokisch@hftl.de, tristan.langenberg@hftl.de, g-pinter@port.kobe-u.ac.jp

Abstract

State-of-the-art pronunciation tutoring (CAPT) systems are based on ASR technology. Consequently, they can provide a distinguished learning feedback which is focused on phonetic features and the positions of articulation errors. In contrast with the relative success with segmental errors, the acquisition and assessment of second language (L2) prosody is still a challenging problem. Although prosodic parameters like f_0 contour or duration measures are usually displayed, the consequential evaluation components are generally missing. Considering the strong variation in speech data, functional data analysis (FDA) is a useful concept which statistically analyses interrelations between principal components (e.g., given accentuation) and their contribution to superimposed forms (e.g., resulting f_0 contour). This article describes baseline processing and preliminary results of a pilot study on the intonation-based proficiency classification of German by using FDA methods. The experimental part contains the FDA-based classification results compared to a perceptual classification by German natives.

Index Terms: L2 prosody, proficiency, functional data analysis

1. Introduction

Computer-assisted language learning (CALL) and so-called intelligent language tutoring systems (ILTS) have been established components in second language education for more than a decade. Among the proposed methods and systems, automatic pronunciation tutoring (CAPT) plays an increasingly important role. Available CAPT systems offer a wide range of user feedback, such as recorded and reference waveforms, analyzed spectra and underlaying phoneme sequence or animated articulatory organs—often including the intonation contour of uttered phrases. Nevertheless, the pronunciation assessment including the marking of error positions is usually based on segmental (phonetic) features and relies on conventional automatic speech recognition (ASR) modules that rely on hidden Markov models (HMMs) and, for example, use Goodness of Pronunciation (GOP) score as confidence measure [1]. In the system development, elaborate speech databases (originally developed for ASR) can be reused. Although the importance of the prosody acquisition is widely agreed among linguists and teachers, research and development have limited focus on suprasegmental (prosodic) evaluation components. This lack of interest might be surprising, since prosodic core parameters like f_0 contour or rhythmic structures can be easily measured. During the development of the CAPT systems AzAR and Euronounce by TU Dresden and partners [2, 3], effort was invested in suprasegmental databases for the assessment of cross-lingual effects in the acquisition of second (L2) or third language (L3) prosody. The Euronounce database contains 130 speakers of German, Polish, Czech, Slovak and Russian (including 18 language students per L1/L2 pair) and about 200 hours of speech. In further projects the AzAR concept and databases were extended to Mandarin learners of German, to L2 learners of Basque [4, 5, 6], and a baseline method to evaluate intonation contours was suggested.

Considering the strong variation in speech, we found that the functional data analysis (FDA) introduced by Ramsay and Silverman [7, 8] can also provide a powerful approach in speech analysis by statistically exploring interrelations between principal components (e.g., accentuation) and their contribution to forms (e.g., f_0 contour). FDA-based methods in prosodic analysis and synthesis have been already suggested by Gubian et al. [9, 10]. In a recent study, Ward [11] applied principal components analysis (PCA) to several dozen contextual prosodic features in a large set of heterogeneous dialog data. The resulting prosodic components are interpretable as prosodic patterns, including some which involve behaviors of both interlocutors. We intend to apply FDA in different stages of the prosodic assessment—focused on CAPT. In the current article, we describe preliminary results of a pilot study on an automatic intonation-based proficiency classification for German language to test the potential of FDA methods in CAPT environment. In our case the proficiency classification by limited (i.e., only intonational) information is just a working assumption. A detailed prosodic analysis of single speaker utterances or a reliable ”overall” proficiency level classification of a speaker is not intended within the scope of this paper. It is clear that the spoken language proficiency is characterized by complex feature sets such as active vocabulary, rules of grammar and phonology, phonetic correctness and so on. Section 2 introduces some previous work on L2 prosody assessment and provides links to proficiency classification. In section 3, we briefly explain the FDA concept. The pilot study on 16 speakers of German is described in section 4—including the L1/L2 database, the baseline processing and the experimental results. The results consist of two parts—addressing FDA-based classification results and a listening test with proficiency classification by native speakers of German.

2. Proficiency and prosody assessment in second language learning

2.1. Proficiency classification

Standardized tests of language proficiency such as the Test of English as a Foreign Language (TOEFL) [12] were already established in the 1960s focusing on reading, listening and writing rather than speaking abilities. In the TOEFL Internet-based Test (iBT) since 2005, the performance evaluation in reading and listening is based on questionnaires. The results of writing and speaking sections are evaluated by three to six human raters which is costly. Consequently, automatic classification methods

10.21437/SpeechProsody.2014-147

are mainly addressing the written proficiency of language learners using algorithms from machine translation [13, 14, 15]. The assessment of spoken language proficiency is focused on phonetic features using GOP or similar measures as already discussed in section 1. Features of non-native prosody, which limit the L2 proficiency, have been studied in different contexts—focused on L2 American or British English. Within the context of this article, some studies dealt with the non-native accent identification [16, 17, 18].

2.2. Prosodic assessment

Beyond the native versus non-natives classification problem, studies of Hönig, Batliner et al. [19, 20] tried to assess L2 productions with respect to intonation and rhythm on a continuous scale, and suggested a suitable set of prosodic features that approximated the decisions of human labelers. In [21], the surveyed feature spaces were extended by acoustic features known from speaker identification tasks (such as short-time spectral features) or general-purpose features from established paralinguistic analysis to indirectly capture complementary prosodic information. The studies consider the perceptual evaluation as a reference—as we do in our work—and describe promising classification strategies. Nevertheless, by fusing different prosodic or even complementary features, the impact of single components (e.g., specific word accentuation or phrase modus variation in the intonation contour) can not be adequately modeled, which results to a less specific user feedback. Our study is targeting on principal intonation components which contribute to the proficiency classification—as an indicator for the influence on L2 prosody—and not on the overall optimal feature representation and classification.

In a previous study [6], we identified Basque as an interesting object of L2 studies in prosody. Basque is an isolated language which does not belong to the Indo-European language group, as one would expect from its geographical location. It has two major neighboring languages, Spanish and French, and the influence of these languages on Basque is noticeable—especially for people studying L2 Basque. In this light it is not surprising that Euskaltzaindia, the Academy of the Basque language, has not yet made a decision about standard prosody due to the variety of accents and intonations across the dialects. L2 students of Basque do not have a clear reference of the preferable pitch pattern and often opt for that of their own L1. This indeterminacy effects the application of conventional quantitative intonation models. The proposition of the Basque study assumes the teacher voice including its intonation pattern as a reference. By providing same text example to the student (cf. ”shadowing task”), the intonation quality is simply indicated by the root mean square error (RMSE) between realized f_0 contour of the student and the according reference utterance. For this purpose, f_0 contours need to be normalized on their mean value (gender normalization) and synchronized to the reference by dynamic time warping (DTW) as shown in figure 1.

3. Functional data analysis in prosody

In general, functional data analysis (FDA) names a branch of statistics that analyzes multivariate data which may be treated preferably as curves or surfaces varying over a continuum which is often time, but it can be, among others, spatial location or probability. The data may be subject to measurement errors or even have only an indirect relationship to the curve that they define. It is assumed that the curves are intrinsically smooth.

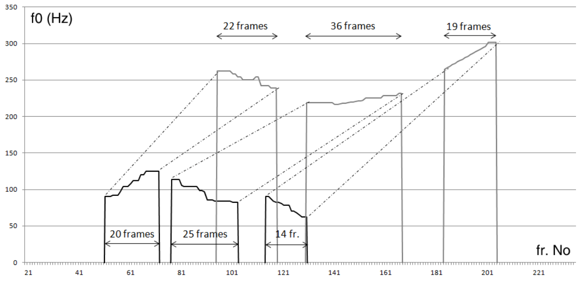

Figure 1: Exemplary f_0 contour mapping: Basque male (reference below) vs. Slovakian female uttering Basque word INDEPENDENTZIA from [6].

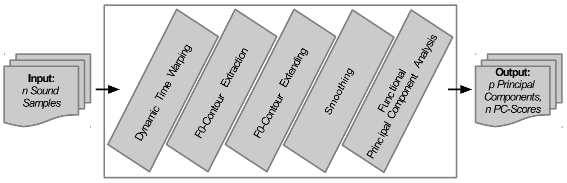

Figure 2: f_0 analysis steps in overview.

Ramsay, Silverman and others developed a set of descriptive and exploratory FDA techniques [7, 8]. Functional data analysis can also use information of slopes and curvatures—reflected in derivatives of the curves—which may reveal aspects of the data generation process. Functional data models and methods resemble those for conventional analyses of multivariate data, including smoothing techniques, regression models, splines and principal components analysis. Gubian et al. introduced the FDA to speech analysis [9] and, in particular, to the f_0 analysis (e.g., for discriminating questions and answers as in [10]). According to Gubian, FDA provides a qualitative/visual description of results and quantitative output in form of statistical values and can be, therefore, interpreted as an “interface between shapes and numbers”. We follow this approach and use functional data analysis to abstract away from random f_0 variation in instances of the same utterance. Moreover, the FDA shall extract a maximum of relevant information from our dataset and supports a reliable estimation of the f_0 curves. The analysis process consists of five steps as shown in figure 2. In the first step, all utterances (waveforms) are aligned by dynamic time warping (DTW) to avoid timing mismatch. The subsequent f_0 curve extraction is based on Praat [22]. In the next steps, unvoiced parts are approximated with splines and the resulting contours are smoothed. The final part, the principal component analysis (PCA) for functional data, is the most important analysis step. The PCA reduces the number of functions in the dataset to a lower number of principal components (PCs). The PCs, based on eigenvalues and eigenfunctions by a complex reduction, represent a maximum of information of the whole dataset. For the PCA execution, we need the covariance functions of our data as described in [8],

$$v(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(s) - \overline{x}_i(s)] \cdot [x_i(t) - \overline{x}_i(t)] \quad (1)$$

and the definition of the extreme values for the eigenvalues

$$\mu = max_{\xi} \left\{ \sum_{i=1}^{n} [\int \xi(t) x_i(t)\, dt]^2 \right\}. \quad (2)$$

Henceforward one can formulate an eigenvalue problem:

$$\int v(s,t) \xi_j(t)\, dt = \mu_j \xi_j(s). \quad (3)$$

This problem is transformed into the form $\mathbf{V}\vec{\xi} = \mu$ and is solved in the numerical linear algebra by the aid of determinants and the unity matrix which results in PCs—built with their eigenfunctions and eigenvalues (cf. [10]),

$$PC_j(t) = \overline{x}(t) + \sqrt{\mu_j} \cdot \xi_j(t) \quad (4)$$

and principal component scores (PC scores) which are important for the subsequent classification (cf. [8]),

$$\mathbf{C}_{scr}(i,j) = \sum_{j=1}^{p} \sum_{i=1}^{n} \int \xi_j(t)[x_i(t) - \overline{x}(t)]\, dt. \quad (5)$$

By analyzing n utterances (in our case study 16 speaker samples), we calculate p principal components (in our case study 3 PCs) and obtain n (p-dimensional) principal component scores (16 PC scores).

4. Pilot study

4.1. Test design and speaker database

The simple test design aims at the question whether the extracted three principal f_0 components of a short test phrase contain relevant prosodic information to solve two practical targets—as intermediate step for the proficiency assessment:

- Speaker discrimination into native or non-native (L1 vs. L2 speaker of German),
- Proficiency classification on a six-point scale (in the style of the language levels A1, A2, B1, B2, C1 and C2 of the Common European Framework of Reference [23]).

The German test phrase NEIN, SIE KANN ES NICHT. ('No, she can not do it'.) was uttered by 16 speakers (one example per speaker). The test dataset includes eight native (L1 German) speakers and eight non-native speakers with mother tongue (L1) Russian. By disregarding further potential language interferences (e.g., L2/L3 English/German), we consider all non-natives as L2 German speakers. Table 1 summarizes the dataset. The reference classification of the speakers to the mentioned language levels is based on the decision of a language teacher. To simplify the task, we consider the highest L2 speaker level C2 equivalent to the mother tongue level of the L1 speakers (level 1 in the six-point scale).

Table 1: *Test database in overview*

Group	Speaker description	No. of speakers
1	L1 German, male ($L1m_x$)	4
2	L1 German, female ($L1f_x$)	4
3	L2 German, male ($L2m_x$)	4
4	L2 German, female ($L2f_x$)	4

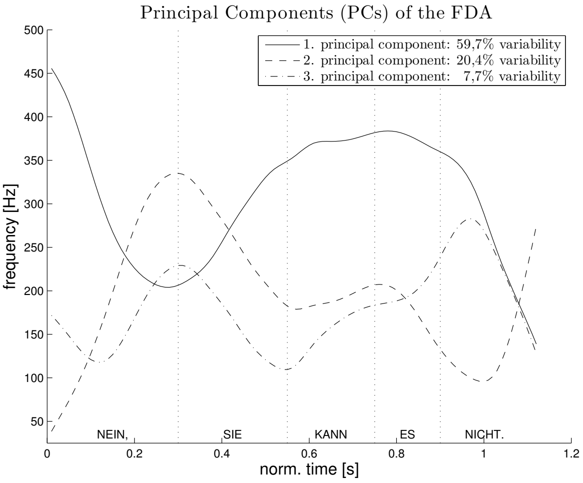

Figure 3: Analyzed principal components of the German phrase: NEIN, SIE KANN ES NICHT. ('No, she can not do it.')

4.2. FDA-based proficiency classification

The first three principal components (PCs) and 16 speaker-based PC scores are calculated as described in section 3. As a reference, the center of mass of each speaker group (L1m, L1f, L2m, L2f) in the PC scores is determined. The root mean square error (RMSE) of each group serves as distance measure. Each speaker distance to the center of mass is divided by the according group RMSE which results to a speaker-specific value of belonging to the certain group—enabling the discrimination into native and nonnatives. By using the PC scores, a geometric distance matrix for the reference speakers is generated (also weighting the PC score dimensions by their information content. The mean deviation of each speaker to its reference value in the matrix is divided by the mean sum of the reference matrix which generates a normalized value of belonging to a certain language level.

4.3. Experimental results

Figure 3 visualizes the first three PCs of the test phrase—representing an information content of 87.8% which we assume as a sufficient accuracy regarding the simple test design. Figure 4 displays all PC scores in two-dimensional PC spaces which leads to three PC combinations with $n = 16$ scores per diagram. The PC scores can be associated with the n f_0 contours. The L1m (male) and L1f (female) speakers create scatter plots which might be associated with feature clusters. Six of the eight L1 speakers can be visually classified. Nevertheless, the L2 speakers do not form visible clusters. Additionally, two L2 speakers seem to cluster with the L1 group.

4.4. Perceptual test

Besides the known proficiency classification provided by language teacher, we performed a perceptual test with 15 naive subjects—ten males and five females with a mean age of 26.3 years. The listeners evaluated 16 utterances (same phrase) in random order and had no prior knowledge about prosodic analysis or the surveyed interrelation to the proficiency assessment.

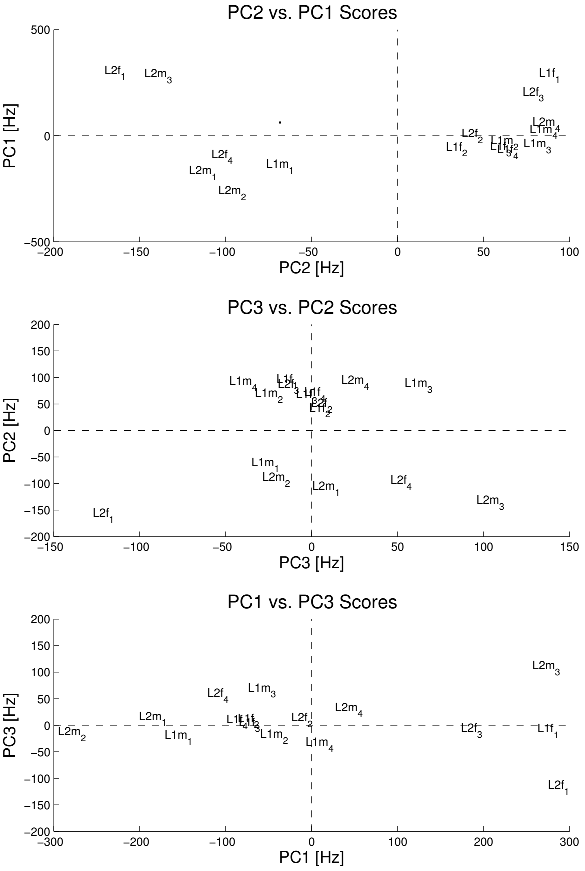

Figure 4: Speaker-specific visualization of the PC scores.

The questionnaire contained the following items:

- Speaker gender (m/f),
- Native speaker of German (y/n),
- Language level on a six-point scale (“1” best).

In L1/L2 classification, the mean decision has a deviation of 0.16 on the normalized scale (“0” L2 . . . “1” L1). In the language level classification, the decisions deviate by 0.98 on the normalized scale (“1” L1 speaker . . . “6” L2 beginner).

4.5. Comparison of FDA-based and perceptual results

By setting the decision threshold to 0.5, the resulting native/non-native classification is shown in figure 5. The “regular classification” is given by the teachers’ reference. The listeners’ classification (“subject group”) is correct for all L1 speakers but includes two errors (speakers 5 and 8) and one border case (speaker 6) in the L2 speaker assessment. The FDA-based classification fails twice in both speaker groups (speakers 1, 5, 12 and 14). Figure 6 visualizes the language level classification of the speakers on the six-point scale. The teachers’ classification is assumed as reference. Both, the mean subjective and the FDA-based decisions are correlated to the reference. For nine speakers, the subjective decision closer reflects the teachers’ classification. In the remaining seven cases, the FDA-based classification is more accurate.

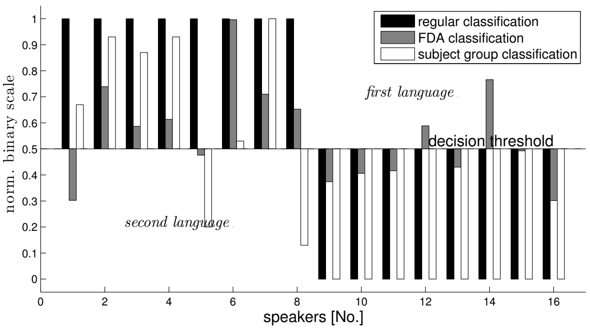

Figure 5: L1/L2 classification by teacher (”regular”), subjects and FDA

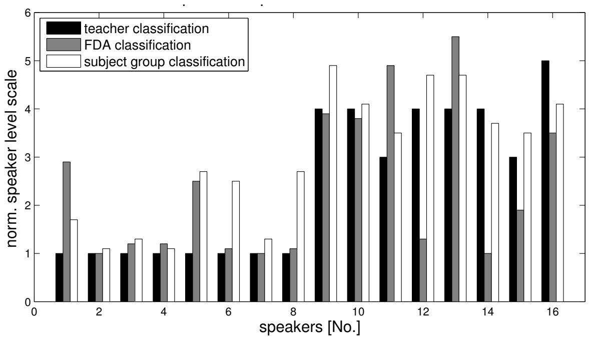

Figure 6: Language level classification by teacher, subjects and intonation-FDA-based (“1” L1 speaker . . . “6” L2 beginner).

5. Discussion

As we expect from our daily-life experience, naive listeners are able to identify non-native speakers and even classify their proficiency on a language level scale with a certain accuracy (widely reflecting the teachers’ assessment). In the test design, listeners can use all noticeable problems in a single test phrase such as mispronounced phonemes, wrong segmental durations or accentuation. In contrast, the FDA-based classification is only leaned on the scoring of three principal components extracted from the f_0 contour. It is remarkable that the simple FDA-based classification, which is only using f_0 information and a highly reduced data description, leads to correct classification in the majority of the L1/l2 speaker decisions. The FDA-based language level classification achieves similar results as the perceptual testing, too.

6. Conclusion

The pilot study shows the potential of the functional data analysis in the proficiency assessment but the results are preliminary and need to be consolidated in further experiments with additional data (i.e., more speakers, phrases and variants, further prosodic parameters). In our study we focused on the question how the FDA concept can be utilized for prosodic analysis in the context of pronunciation training. As a preliminary study, two hypothetic classification tasks were carried out with a few simplifications, and only linear classifiers were used. Further extensions of this study will use trainable classifiers and investigate the viability of these methods in real-world assessments in pronunciation tutoring.

7. Acknowledgements

The L1/L2 test data are part of the Euronounce corpus and were recorded by Rainer Jäckel, TU Dresden. We would also like to thank Michael Graf and Ines Rennert, Leipzig University of Telecommunication (HfTL), for their valuable advice.

8. References

[1] Witt, S. M. and Young, S., “Phone-level pronunciation scoring and assessment for interactive language learning,” J. Speech Communication, vol. 30, 95–108, 2000.

[2] Jokisch, O., Koloska, U., Hirschfeld, D., Hoffmann, R., “Pronunciation learning and foreign accent reduction by an audiovisual feedback system,” in proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII), Beijing, 419–425, 2005. Springer LNCS-3784.

[3] Jokisch, O., Jäckel, R., Rusko, M., Demenko, G., Cylwik, N., Ronzhin, A., Hirschfeld, D., Koloska, U., Hanisch, L., Hoffmann, R., “The EURONOUNCE project - An intelligent language tutoring system with multimodal feedback functions: Roadmap and specification,” in proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 116–123, 2008. Frankfurt.

[4] Ding, H., Jokisch, O., Hoffmann, R., “F0 analysis of Chinese accented German speech,” in proc. 5th Intern. Symposium on Chinese Spoken Language Processing (ISCSLP), 49–56, 2006. Singapore.

[5] Hilbert, A., Mixdorff, H., Ding, H., Pfitzinger, H., Jokisch, O., “Prosodic analysis of German produced by Russian and Chinese learners,” in proc. 5th Intern. Conf. on Speech Prosody, 2010. Chicago.

[6] Odriozola, I., Jokisch, O., Hernaez, I., Hoffmann, R., “A Pronunciation Tutoring System for Basque - First Development Steps,” in proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 2012. Cottbus.

[7] Ramsay, J. O. and Silverman, B. W., “Functional Data Analysis”, Springer, New York, 1997.

[8] Ramsay, J. O., Hooker, G. and Graves, S., “Functional Data Analysis with R and MATLAB”, Springer, 100–103, New York, 2009.

[9] Gubian, M., Torreira, F., Strik, H., Boves, L., “Functional data analysis as a tool for analyzing speech dynamics – a case study on the French word c’´etait,” in proc. Interspeech, 2199–2202, Brighton, 2009.

[10] Gubian, M., Boves, L. and Cangemi, F., “Joint analysis of f_0 and speech rate with Functional Data Analysis,” in proc. ICASSP, 4972–4975, Florence, 2011.

[11] Ward, N. G., “Automatic discovery of simply-composable prosodic elements,” in proc. 7th Intern. Conf. on Speech Prosody, Dublin, 2014.

[12] Educational Testing Service (ETS), “Test of English as a Foreign Language (TOEFL)”. Retrieved December 1, 2013 from https://www.ets.org/toefl

[13] Corsten-Oliver, S., Gamon, M. and Brockett, C., “A machine learning approach to the automatic evaluation of machine translation,” in proc. 39th Annual Meeting of the Association for Computational Linguistics (ACL), 148–155, Toulouse, France, 2001.

[14] Lee, J., Zhou, M., Liu, X., “Detection of non-native sentences using machine-translated training data,” in proc. Human Language Technology Conference of the North American Chapter of the the Association for Computational Linguistics (NAACL HLT), 93–96, Rochester NY, 2007.

[15] Kotani, K., Yoshimi, T., Kutsumi, T. and Sata, I., “Automatic classification of language learner sentences into native-like or non-native-like based on word alignment distribution,” in W. Kouwenhoven (Ed.): Advances in Technology, Education and Development. InTech, Rijeka, 2009.

[16] Tepperman, J., Narayanan, S., “Better non-native intonation scores through prosodic theory,” in proc. Interspeech, 1813–1816, Brisbane, 2008.

[17] Piat, M., Fohr, D. and Illina, I., “Foreign accent identification based on prosodic parameters,” in proc. Interspeech, 759–762, Brisbane, 2008.

[18] Lopes, J., Trancoso, I. and Abad, A., “A nativeness classifier for ted talks,” in proc. ICASSP, 5672–5675, Prague, 2011.

[19] Hönig, F., Batliner, A., Weilhammer, K. and Nöth, E., “Automatic assessment of non-native prosody for English as L2,” in proc. Speech Prosody, Chicago, 2010.

[20] Hönig, F., Batliner, A. and Nöth, E., “How many labellers revisited – naives, experts and real experts,” in proc. SLATE, Venice, Italy, 2011.

[21] Hönig, F., Bocklet, T., Riedhammer, K., Batliner, A. and Nöth, E., “The automatic assessment of non-native Prosody: combining classical prosodic analysis with acoustic modelling,” in proc. Interspeech, Portland, Oregon, 2011.

[22] Boersma, P. and Weenink, D., “Praat: Doing phonetics by computer” (version 5.3.05). Retrieved February 24, 2012 from https://www.praat.org

[23] Council of Europe, “Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)”. Retrieved December 1, 2013 from https://www.coe.int/lang-cefr