Phrase Accentuation Verification and Phonetic Variation Measurement for the Degree of Nativeness Sub-Challenge

Claude Montacié [1], *Marie-José Caraty*[1-2]

[1] STIH Laboratory, Paris Sorbonne University, 28 rue Serpente, 75006, Paris, France
[2] Paris Descartes University, 45 rue des Saints-Pères, 75006, Paris, France
Claude.Montacie@paris-sorbonne.fr, Marie-Jose.Caraty@ParisDescartes.fr

10.21437/Interspeech.2015-180

Abstract

The Degree of Nativeness Sub-Challenge consists in the automatic grading of the pronunciation quality of non-native English utterances. In this paper, we investigate the phrase accentuation and the phonetic acoustic variability for the prediction of the grades. Two prediction systems have been developed: the Extended Baseline System (EBS) and the Pronunciation Feature based System (PFS). The EBS system was designed to take into account the cross-corpus specificities such as recording conditions and the sentence variability. The speech files were segmented using Automatic Speech Recognition methods (ASR). Audio features were selected on both the training and development sets using the Regressional ReliefF method. New audio features were developed for the PFS system to take into account the mispronunciations: unusual prosody and/or phonetic variation. These systems have been assessed using the Spearman's correlation coefficient with expert annotations. The PFS system has significantly improved of 0.05 the Official Baseline performance.

Index Terms: pronunciation grading, phrase accentuation, accent classifier, phonetic variation, computational paralinguistics

1. Introduction

The Degree of Nativeness (DN) Sub-Challenge [1] consists in the automatic assessment of the pronunciation quality of non-native utterances. A foreign accent can lead to difficulties with speech intelligibility and comprehensibility [2, 3]. Various degrees of foreign accent can be perceived [4, 5]. The assessment of non-native speech is central to language teaching. One of the first difficulties for learners of a second language (L2) is the mispronunciation of L2 phonemes [6]. Many studies have investigated the weakness in discriminating between the acoustic variability of an L2 phoneme and the closer phonemes in the L2 phonemic system [7]. This lack of discrimination is observable on both perception and production. Such mispronunciations can be explained by the influence of the mother tongue (L1) on the learning process of the L2 [8, 9]. The prosody is the other important factor of the perception of foreign accent. The prosodic errors have an influence on the speech comprehensibility. The accentuation is known to speed up the comprehension by focusing on the new information [10]. As for the acquisition of phonological skills, many studies are published on the transfer from L1 to L2 prosody acquisition [11, 12].

The automatic assessment of pronunciation quality is an important part in the Computer-Aided Language Learning (CALL) systems. The majority of studies were conducted on the detection of errors at the phone level. The non-native utterances were segmented using ASR systems based on Hidden Markov Models (HMM) [13] trained on native utterances. An estimation of the quality of a phone pronunciation can be given by the posterior probability of this phone [14]. The score of pronunciation quality of the utterance is the combination of the posterior probabilities. Various methods have been developed to normalize the posterior probabilities: the calibration using neural networks [15, 16], the Goodness Of Pronunciation algorithm (GOP) [17], the posterior probabilities given by a forced alignment [18] or by an ASR system using HHM trained on non-native utterances [19]. The prosodic quality is estimated using statistics of the phone duration [20, 21, 22] such as the vocalic duration ratio. Other approaches have been studied such as the spectral analysis of specific phonemes [23] or the assessment of melodic information [24, 25].

The cross-corpus setting is the specificity of the DN Sub-Challenge [1]. However, in the previous systems, the HMM were trained using the same guidelines of collecting speech for native and non-native utterances. For the challenge, our hypothesis is that the approaches based on the posterior probabilities can be sensitive to the cross-corpus setting. We paid a particular attention to the acoustic features for the DN assessment using only non-native speech material. We have studied the baseline audio feature set of the challenge looking for a feature representation well fitted to the prosody verification of acceptable phrasing, and to allophonic variation.

The paper is organized as follows. In Section 2, the speech corpora of the DN Sub-Challenge and the Official Baseline System (OBS) are described. In Section 3, the OBS system is extended to take into account the cross-corpus specificities. Three preprocessing methods were developed for the Extended Baseline System (EBS) for the normalization of the various corpora: (1) segmentation of the speech files using ASR-based segmentation, (2) filtering of the training corpus for the adequation to the other corpora, (3) a joint selection of audio features over the corpora. In Section 4, a Pronunciation Feature based System (PFS) is described. New audio features were designed to take into account the mispronunciations: unusual prosody and phonetic variation. The last section concludes the study.

2. Speech material and Baseline system

Four English language databases of non-native speakers were used to build up the training (Train), development (Devel) and test (Test) sets of the DN Sub-Challenge. The training set was made of a subset of the AUWL [26] and ISLE [27] databases. The development set was made of a subset of the C-AuDiT database [28]. The test set was created and recorded at TUM [1]. The Train, Devel and Test sets are speaker- and sentence-independent. The specificities of the cross-corpus setting of the DN sub-challenge are described below.

Table 1. *Statistics on the speech material.*

Corpora	Train	Devel	Test
# of speech files	3,890	999	594
Recording conditions	very heterogeneous	heterogeneous	homogeneous
Speaking style	prepared	read aloud	read aloud
# of speakers (female, male)	67 (24, 43)	58 (31, 27)	54 (28, 26)
# of L1 language	8	5	unknown
Duration (sec) Average: min-max	5.3: 1.8-23.5	9.7: 3.8-28.8	8.5: 2.9-22.4
# of words Average: min-max	9: 1-38	15: 4-45	19: 8-36
# of phonemes Average: min-max	29: 3-141	51: 17-148	63: 26-111

Table 1 gives for each corpus some characteristics and statistics on the Train, Devel and Test sets. The number of words is given per speech file representing a sentence. The number of phonemes per sentence is obtained from a forced alignment using an ASR system. We note that the number of phonemes in average is 29 for Train, 51 for Devel and 63 for Test. The shortest sentence contains 3 phonemes for Train, 17 for Devel and 26 for Test. These characteristics show the difficulties of the DN assessment within the cross-corpus scenario.

For the speech files of each set, two kinds of metadata are also provided: the word sequence to be produced and the acceptable phrasing in terms of phrase accents (primary, secondary, none) and boundaries (major, minor). Specifically for the Train and Devel sets, two additional metadata are supplied: the DN score, the automatic phoneme segmentation and the speaker identity. The DN score was obtained by expert annotations according to sentence melody and rhythm. A phonetic dictionary of the words occurring in the corpora is also furnished. The official performance of DN prediction system uses the Spearman correlation coefficient (ρ) with the DN scores.

The OBS system [1] is as follows. The baseline audio feature set (6373 features) [29, 30] is used to describe the speech files in terms of spectral, cepstral, prosodic and voice quality information. Linear Support Vector Regression (SVR) [31] is used for the DN prediction. SVR is based on Support Vector Machines (SVM) with linear kernel. The Sequential Minimal Optimisation (SMO) algorithm was used for training. The insensitive-loss SVR parameter was chosen to 1.0. The SVM complexity parameters were 10^{-5} for the DN prediction on the Devel set and 10^{-4} on the Test set. The Baseline performances in terms of ρ coefficient are 0.415 on the Devel set and 0.425 on the Test set.

3. Extended Baseline system

Taking into account the variety of the corpora, preprocessing methods of normalization were developed for the Extended Baseline System (EBS). Three major differences between corpora have been observed: the speech file segmentation, the sentence size in phoneme number and the conditions of collecting speech. The SVR machine described in the previous section was used for the DN prediction.

3.1. Segmentation of the speech files

The duration of non-speech is very variable before and after the speech signal. Non-speech includes silence, noise and speaker sounds such as breath, cough and lip smack. The speech segmentation is based on an acoustic-phonetic decoding system followed by an automatic localization of the speech boundaries. The version 0.8 of the Pocketsphinx recognizer library [32] was used to develop the ASR system. The acoustic models were the generic US-English acoustic models provided by CMU [33]. A significant improvement (+0.083) of the ρ coefficient is obtained compared to the Baseline performance on the Devel set (0.415). This improvement is confirmed on the Test set (see Section 5). This segmentation pre-processing was used for the next experiments on the Train, Devel and Test sets.

3.2. Filtering of the training set

On the assumption that the shortest sentences in terms of phonemes (e.g., "thanks", "exactly", "you're welcome") are not useful for the DN prediction of longer sentences, experiments were carried out on the effect of the filtering of short sentences. Using the previous segmentation of the speech files, the number of phonemes was obtained by a forced alignment. The HTK toolkit [34] was used for this alignment and the acoustic models were trained again. A new Train set was built up by removing the speech files containing a number of phonemes inferior to a given threshold. In Figure 1, the C2 curve shows the number of speech files of the new Train set as a function of a given threshold. The C1 curve shows the ρ coefficient of the DN prediction on the Devel set using the new Train set.

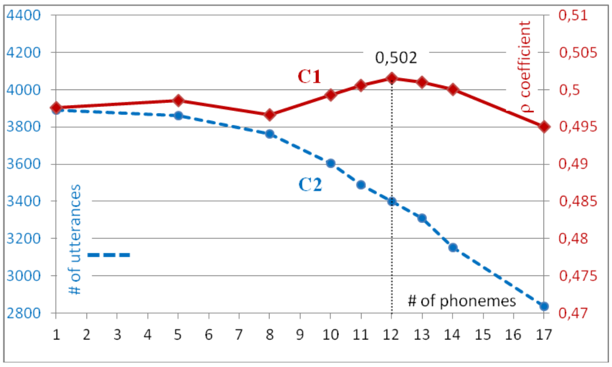

Figure 1. *C1, the ρ coefficient of the DN prediction on Devel using the new Train, C2, the new Train size as a function of a minimum number of phonemes per speech file.*

A not significant improvement (+0.004) of the ρ coefficient is obtained with a threshold of 12 phonemes. The corresponding training set is made of the 3,398 speech files of the Train set. This set is used as Train set for the next experiments. The benefit of this pre-processing is confirmed on the performance of the DN prediction on the Test set (see Section 5).

3.3. Joint feature selection

In the cross-corpus setting, the methods of speech collection vary: shadowing with a reference speaker, repeating after the reference speaker or reading aloud a prompted text. The linguistic characteristics of the speech data can vary in terms of accentuation, rhythm and phonetic realizations. Some audio features can be relevant in the DN prediction using a collection method and irrelevant for another one. The joint selection method consists in keeping the N most relevant features of the Train with N maximizing the ρ coefficient on the Devel set. The Regressional ReliefF (RRF) [35] method was used to rank the audio features.

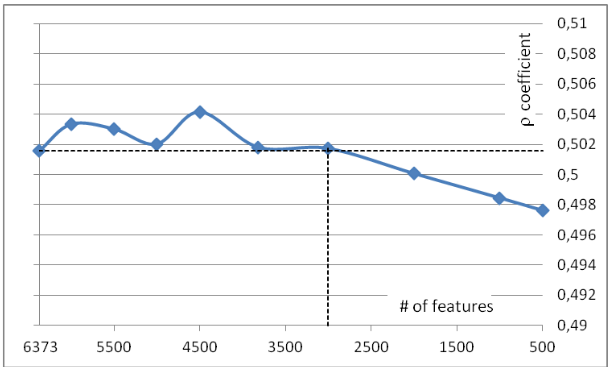

Figure 2. *ρ coefficient of the DN prediction on Devel according to the number of features selected on Train.*

Figure 2 gives the ρ coefficient on the Devel set as a function of the number of selected features. We can note that it is possible to remove 50% of the features set without any loss in performance. A not significant improvement (0.002) of the ρ coefficient is obtained for a joint selection of 4,500 features.

4. Pronunciation features

To take account of the DN prediction task, 18 new audio features were designed to fit the prosody verification, and the allophonic variation in order to detect the mispronunciations. Three types of features have been computed: statistics on speech temporalities (10 features), dissimilarity measure between the detected accents and an acceptable phrasing (2 features), and the spectral variance of allophonic variants (6 features). Each new audio feature has been assessed by the RRF rank and its Spearman's correlation with DN. If the ρ coefficient is positive, the audio feature varies as DN (+), else in inverse ratio (–). The RRF rank was computed on 6,391 features (18 new features added to the 6,373 baseline features).

4.1. Speech temporality features

In related works, various speech temporalities features have been developed [21, 36, 37]: speech rate measures, global interval proportions and pairwise variability index. We chose to use four speech rate measures and six new features related to the boundaries of acceptable phrasing. The speech rate measures are Speaking rate (S), Articulation rate, (A), Pause duration ratio (Pd) and Pause occurrence ratio (Po) using the phonetic segmentation of the speech files. The six new features are computed using an alignment between an acceptable phrasing provided in the prompts (boundaries and accentuations) with the phonetic segmentation.

- Boundary pause duration ratio (Bd): the total duration of pauses divided by the number of boundaries.
- Boundary pause occurrence ratio (Bo): the number of pauses divided by the number of boundaries.
- Major boundary pause duration ratio (Md): the mean duration of pauses corresponding to a major boundary.
- Major boundary pause occurrence ratio (Mo): the number of pauses corresponding to a major boundary divided by the number of major boundaries.
- Minor boundary pause duration ratio (md): the mean duration of pauses corresponding to a minor boundary.
- Minor boundary pause occurrence ratio (mo): the number of pauses corresponding to a minor boundary divided by the number of minor boundaries.

Table 3. *Speech temporality feature ranks (on 6,391) and sign of the ρ coefficient on Train and Devel.*

Feature	Feature rank & ρ sign (+ / –)		Feature	Feature rank & ρ sign (+ / –)	
	Train	Devel		Train	Devel
S	624 –	15 –	A	1,014 –	48 –
Pd	29 +	9 +	Po	1 +	12 +
Bd	19 +	326 +	Bo	108 +	31 +
Md	5 –	6,272 –	Mo	51 –	6,312 –
md	391 –	15 +	mo	685 –	32 +

In Table 3, known results on the speech temporalities of non native utterances are confirmed. The speaking and articulation rates vary in inverse ratio to DN score(–). Most of the new features are relevant for DN prediction on both the Train and Devel sets. The two exceptions are the Md and Mo features on the Devel set. One explanation is the difference of the methods of speech collecting between the Train and Devel sets.

4.2. Accentuation features

The accentuation is an important factor of the perception of foreign accent. Our assumption is that the proposed phrasing will be used by the reference speakers and imitated by the non-native speakers. We chose to develop an accent classifier to verify this assumption. Two classes were considered: primary accent (P) and secondary accent (S). Two corpora were built up from the speech files annotated with the best DN score. These corpora were used to train and assess the accent classifier. The accent training set consisted in 1,067 words (663 with a P accent and 404 with an S accent) extracted from the Train set. The accent development set consisted in 227 words (115 with a P accent and 112 with an S accent) extracted from the Devel set. The baseline audio feature set (6,373 features) was used to describe the speech files. SVM using logistic regression allows the computation of a posterior probability of the accentuation labels (P or S) for each word. The performance of the classifier is equal to 65% in unweighted average recall on the accent development set. Two features were computed from the posterior probabilities of the words of a speech file: the arithmetic (Am) and geometric (Gm) mean of the posteriors of the accent classifier.

Table 4. *Accentuation feature ranks (on 6,391) and sign of the ρ coefficient on Train and Devel.*

Feature	Feature rank & ρ sign (+ / –)		Feature	Feature rank & ρ sign (+ / –)	
	Train	Devel		Train	Devel
Am	9 –	6,305 –	Gm	130 –	6,090 –

In Table 4, the accentuation features vary in inverse ratio (–) to DN. We can note that these features are relevant for the Train set and irrelevant for the Devel set. One explanation may be that the speakers of the Devel set can have used another acceptable phrasing than the reference one, for which the Train speakers were prepared. An improvement of the accentuation feature could be the generation of the various acceptable phrasings.

4.3. Allophonic features

Approximation in the realization of phones can be a characteristic of mispronunciation. Audio feature based on the variance of the phone realizations were developed. Phone realizations are represented by the MFCC vector in the middle of the phonetic segment. Four phonetic classes have been selected according to their phonetic proximity and the RRF rank on the Train set: {I}, {i, i:}, {I, i, i:}, {p, t, k}.

Table 5. *Allophonic feature ranks (on 6,391) and sign of the* ρ *coefficient on the Train and Devel sets.*

Feature	Feature rank & ρ sign (+ / –)		Feature	Feature rank & ρ sign (+ / –)	
	Train	Devel		Train	Devel
{I}	1,118 –	1,500 +	{i, i:}	1,898 –	336 +
{I, i, i:}	1,041 –	718 +	{p, t, k}	1,386 –	764 +
{d, t}	534 –	2,232 +	{k, g}	2,812 –	82 +

In Table 5, the allophonic features vary in inverse ratio (–) to DN score for the Train set. This is an unexpected result. The variance of the phones realization should vary as DN score. An explanation is a limitation of the allophonic variation caused by the preparation before recording as the shadowing method of collect.

4.4. Experiments

The PFS system used 6,391 features (18 Pronunciation features + 6,373 baseline features) and the same preprocessing methods as the EBS system. The joint selection method was used to choose the best feature set from the 6,391 features. Figure 3 gives the ρ coefficient on the Devel set as a function of the number of selected features. An improvement (+0.015) of the ρ coefficient is obtained with no feature selection compared to the performance of the EBS system (0.502, see Figure 1). A significant improvement (+0.05) of the ρ coefficient is obtained for a joint selection of 500 features (9 Pronunciation features + 491 baseline features) compared to the performance without feature selection (0.517). The chosen Pronunciation features are {Pd, Bd, Md, md, Po, Bo, Mo, Am, Gm}.

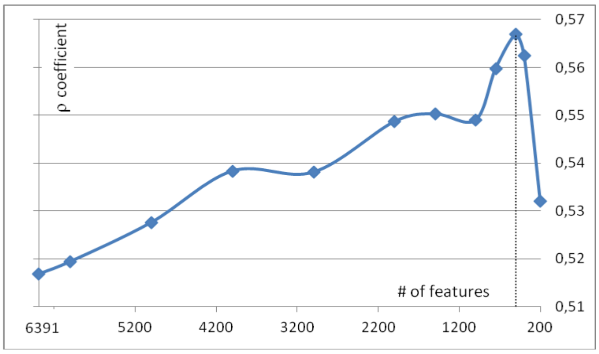

Figure 3. *ρ coefficient of the DN prediction on Devel according to the number of features selected on Train.*

5. Test results

For the submissions on the Test set, the speech segmentation was used for the EBS and PFS systems. Contrastive experiments have been conducted on the joint feature selection and on the threshold used for the Train filtering. Table 6 presents the performance achieved by the EBS and PFS systems on the Test set.

Table 6. *Performances of the EBS and PFS systems on the Test set.*

System	Contrastive conditions	ρ coefficient
OBS	none	0.425
EBS	Train filtering (12 phonemes)	0.457
PFS#1	No train filtering	0.462
PFS#2	Train filtering (12 phonemes)	0.473
PFS#3	Train filtering (13 phonemes)	**0.475**
PFS#4	Train filtering (14 phonemes)	0.470
PFS#5	Train filtering (12 phonemes) Joint selection (1000 features)	0.420

The results showed significant improvements of the ρ coefficient for the EBS (+0.032) and PFS#3 (+0.05) systems compared to the baseline performance (OBS). The contrastive experiments {EBS, PFS#2} and {PFS#1, PFS#3} have shown respectively the contribution of the Pronunciation features (+0.016) and that of the train filtering (+0.013). The joint features selection (PFS#5) did not give the expected results.

6. Conclusion

In this paper, we have presented two DN prediction systems. The first system experiments preprocessing methods to take account of the variety of the corpora. The second system was based on features related to the mispronunciations of non-native speakers. Eighteen audio features were designed: ten speech temporality features, two accentuation features and four allophonic features. A ρ correlation of 0.475 has been obtained on the Test set compared to the Baseline performance of 0.425. However, these performances are not sufficient for the integration of such DN prediction in CALL systems. Future works should include features related to the dissimilarities with the pronunciations of native speakers. The pronunciations of short messages seem to be specific and should be studied for their DN prediction.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in Proc. of INTERSPEECH 2015, ISCA, Dresden, Germany, 2015.

[2] M. Munro, and T. Derwing, "Foreign accent, comprehensibility, and intelligibility in non-native speech," Language Learning, 45, pp. 73-97, 1995.

[3] C. Theodoropulos, "Prosody in the Production and Processing of L2 Spoken Language and Implications for Assessment. Teachers College," Columbia University Working Papers in TESOL & Applied Linguistics, 14(2), pp. 1-20, 2014.

[4] J. E. Flege, and K. L. Fletcher, "Talker and listener effects on degree of perceived foreign accent," The Journal of the Acoustical Society of America, 91(1), pp. 370-389, 1992.

[5] M. G Busà, "On the production of English vowels by Italian speakers with different degrees of accent," in J. Leather and A. James (Eds), New Sounds '92, pp. 47-63, 1992.

[6] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin,. Teaching pronunciation: A reference for teachers of English to speakers of other languages. Cambridge University Press, 1996.

[7] C. Koniaris, G. Salvi, and O. Engwall, "On mispronunciation analysis of individual foreign speakers using auditory periphery models," Speech Communication, 55(5), pp. 691-706, 2013.

[8] T. Odlin, Language transfer: Cross-linguistic influence in language learning. Cambridge University Press, 1989.

[9] S. Curtin, H. Goad, and J. V. Pater, "Phonological transfer and levels of representation: the perceptual acquisition of Thai voice and aspiration by English and French speakers," Second Language Research, 14(4), pp. 389-405, 1998.

[10] Terken, J., & Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. Language and Cognitive processes, 2(3-4), 145-163.

[11] S. Tahta, M. Wood, and K. Loewenthal, "Foreign accents: Factors relating to transfer of accent from the first language to a second language," Language and Speech, 24(3), pp. 265-272, 1981.

[12] L. Rasier, and P. Hiligsmann, "Prosodic transfer from L1 to L2. Theoretical and methodological issues," Nouveaux cahiers de linguistique française, 28, pp. 41-66, 2007.

[13] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in English pronunciation," in ICSLP, Vol. 90, pp. 1185-1188, 1990.

[14] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," Speech communication, 30(2), pp. 83-93, 2000.

[15] H. Franco, and L. Neumeyer, "Calibration of machine scores for pronunciation grading". In ICSLP, pp 2631-2634, 1998.

[16] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers," Speech Communication, 2015.

[17] S. M. Witt, and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech communication, 30(2), pp. 95-108, 2000.

[18] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech". In ICSLP, pp. 1457-1460, 1996.

[19] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning". In EUROSPEECH, pp. 851-854, 1999.

[20] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in Eurospeech, Rhodes, Greece, 1997.

[21] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for english as l2," in Proc. Speech Prosody, Chicago, 2010.

[22] V. Cardeñoso-Payo, C.González-Ferreras, and D. Escudero-Mancebo, "Assessment of Non-native Prosody for Spanish as L2 using quantitative scores and perceptual evaluation," in Proc. of LREC, pp. 3967-3972, 2014.

[23] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," Speech Communication, 51(10), pp. 845-852, 2009.

[24] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez, "Evaluation of speaker's degree of nativeness using text-independent prosodic features," in Proc. of the Workshop on Multilingual Speech and Language Processing, 2001.

[25] A. Rosenberg, "Modeling prosodic sequences with k-means and dirichlet process GMMs," in INTERSPEECH, pp. 520-524, 2013.

[26] 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 (IS ADEPT), Stockholm, Sweden, pp. 21–30, 2012.

[27] 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. of LREC, Athens, Greece, pp. 957–964, 2000.

[28] 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, Wrox all Abbey, 2009.

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

[30] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in Proc. of ACM MM, Barcelona, Spain, pp. 835–838, 2013.

[31] 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, vol. 11, no. 1, pp. 10–18, 2009.

[32] A. Chan, E. Gouva, R. Singh, M. Ravishankar, R. Rosenfeld, Y. Sun, D. Huggins-Daines, M. Seltzer, "The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources," www.cs.cmu.edu/~archan/share/sphinxDoc.pdf

[33] R. Weid, "The CMU pronunciation dictionary", release 0.6, www.speech.cs.cmu.edu/cgi-bin/cmudict, 1998.

[34] G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, and P. Woodland, "The HTK book". Cambridge: Entropic Cambridge Research Laboratory, 1997.

[35] M. Robnik-Sikonja, and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," in: Fourteenth International Conference on Machine Learning, pp. 296-304, 1997.

[36] V. Dellwo, and P. Wagner, "Relationships between rhythm and speech rate", In ICPhS, pp 471–474, 2003.

[37] E. Grabe and E.L. Low, "Durational Variability in Speech and the Rhythm Class Hypothesis". In Laboratory Phonology, pp.515-546, 2002.