# The Automatic Assessment of Non-native Prosody: Combining Classical Prosodic Analysis with Acoustic Modelling

*Florian Hönig*[1], *Tobias Bocklet*[1,2], *Korbinian Riedhammer*[1], *Anton Batliner*[1], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany
[2]Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Germany
hoenig@informatik.uni-erlangen.de

## Abstract

In earlier studies, we employed a large prosodic feature vector to assess the quality of L2 learner's utterances with respect to sentence melody and rhythm. In this paper, we combine these features with two standard approaches in paralinguistic analysis: (1) features derived from a Gaussian Mixture Model used as Universal Background Model (GMM-UBM), and (2) openSMILE, an open-source toolkit for extracting acoustic features. We evaluate our approach with English speech from 94 non-native speakers perceptually scored by 62 native labellers. GMM-UBM or openSMILE modelling alone yields lower performance than our prosodic feature vector; however, adding information from the GMM-UBM modelling or openSMILE by late fusion improves results.

**Index Terms**: computer-assisted language learning, non-native prosody, rhythm, automatic assessment

## 1. Introduction

Non-native prosodic traits limit proficiency in a second language (L2). Prosodic phenomena, located on word level and above, encompass word accent position, syntactic-prosodic boundaries, and rhythm, and help listeners to structure the speech signal and to process segmental, syntactic, and semantic content. Non-native prosodic traits are therefore not mere idiosyncrasies, but often seriously hamper mutual understanding. Thus, they have to be modelled in computer-assisted pronunciation training (CAPT).

A few studies deal with non-native accent identification using prosodic parameters [1–3]. In recent basic research, suprasegmental native traits have been investigated when trying to model language-specific rhythm [4, 5]. Maybe the most important general (prosodic) factor to be modelled in CAPT is non-native rhythm: the English prosody of, e. g. French, Spanish, or Hindi native speakers can sound 'strange' and these speakers are sometimes difficult to understand. We therefore set ourselves the task of automatically assessing the quality of L2 learner's productions with respect to *sentence melody* and *rhythm* on a *continuous scale*. Thus, our approach stands out from studies that just deal with the binary classification problem native vs. non-native. Employing perceptual evaluations as ground truth, we studied the impact of number and type [6] of labellers, and computation of suitable prosodic features [7].

The present paper is motivated by the question: 'Can we improve performance by incorporating other features which are less directly related to prosody, but nevertheless very successful in related areas such as emotion identification?' Our assumption is that the prosodic properties we are trying to assess are reflected in other properties of the speech signal, which might be easier to extract or more robust. A combination of these features might add complementary information and improve accuracy. We compare three kinds of features:

(1) Purely prosodically motivated features based on [7]. The modelling is based on the spoken words and the syllabic and phonetic structure. Using the segmentation of a speech recognizer, different prosodic properties of the segmented units are measured.

(2) Purely acoustic features capturing the distribution of short-time spectral features (Universal Background Model, UBM) with the help of a Gaussian Mixture Model (GMM). In combination with Support Vector Machines, this is a well-known approach in the field of speaker identification [8]. Even if not capturing prosodic information directly, this approach might be able to model other properties of speech that are correlated to prosodic properties. The approach has been shown to deliver competitive results in related tasks such as Emotion Identification [9].

(3) openSMILE [10], a toolkit for computing general-purpose acoustic and prosodic features. It is an established standard for paralinguistic tasks (e. g. [11]). and thus a promising candidate for our aims. We employ it in the usual way, i. e. with a flat analysis structure modelling the acoustics of the whole utterance, without additional information coming from transcription or ASR.

## 2. Material and Human Assessment

The data used in this paper is a combination of two datasets: the English database from our German research projects C-AuDiT [7] and the ISLE database [12]. Our database contains English speech from 58 L2 speakers: 26 German, 10 French, 10 Spanish, 10 Italian and 2 Hindi speakers, and additionally 11 native American English (AE) 'reference' speakers. The ISLE corpus contains recordings of non-native English from German and Italian speakers. When designing our recordings, we took 30 sentences from the ISLE database. From this intersection, three experienced labellers chose five sentences judged as 'prosodically most error-prone for L2 speakers of English' [7].

Taking only speakers that spoke all 5 sentences, we arrived at approx. one hour of speech from 94 speakers. By a web-based perception experiment, 22 native American English (AE), 19 native British English (BE), and 21 native Scottish English (SE) speakers with normal hearing abilities judged each sentence regarding different criteria, answering the following questions on a 5-point Likert-scale (1 is best and 5 worst; for details see [7]):

***int***: DID YOU UNDERSTAND WHAT THE SPEAKER SAID?
***acc***: DID YOU HEAR A FOREIGN, NON-ENGLISH ACCENT?
***mel***: HOW DID THIS SENTENCE'S MELODY SOUNDS?

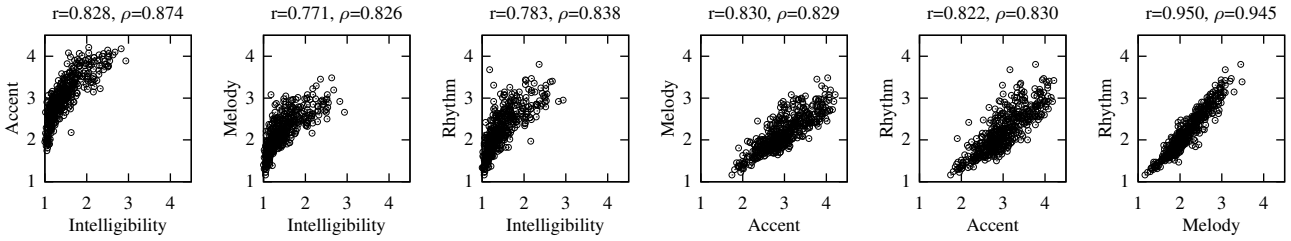


Figure 1: Scatterplots for the annotated scores. For each pair of scores, Pearson correlation $r$ and Spearman correlation $\rho$ is given.

***rhy*:** THE ENGLISH LANGUAGE HAS A CHARACTERISTIC RHYTHM (TIMING OF THE SYLLABLES). HOW DO YOU ASSESS THE RHYTHM OF THIS SENTENCE?

There was no significant difference between the labels of AE, BE and SE listeners. To get a single score for each utterance, we averaged the annotations on the Likert scales across all 62 labellers. Although the correlation between a pair of individual labellers is low, by that, we obtain very reliable annotations [6]. Figure 1 shows the distribution of resulting values and correlation coefficients between the criteria. All four criteria are highly correlated; this can be explained by the observation that segmental and prosodic proficiency are closely related for most L2 learners. Concerning the different ranges on the 1 to 5 Likert scales, we can speculate that the difference between *acc* and *mel*/*rhy* is due to additional segmental errors, and that the lower (i. e., better) *int* value might be traced back to the listeners' language model which is not fully impaired by segmental or suprasegmental errors.

Although we are primarily interested in the prosodic criteria *mel* and *rhy*, we report results for *int* and *acc*, too, firstly, because the labels are highly correlated, and secondly, because it is interesting in itself to relate results for the first with results for the latter.

# 3. Features

## 3.1. Prosodic Features (*pros*)

In order to obtain suitable input parameters for an automatic prosody assessment system, we compute a prosodic 'fingerprint' of each utterance: First, the recordings are segmented by forced alignment of the target utterance using a cross-word triphone HMM speech recognition system. Then, various features measuring different prosodic traits are calculated. They are an extension to those described in [7] and adapted to utterance level instead of speaker level.

We first apply our comprehensive general-purpose prosody module [13] which has proven suitable for various tasks such as phrase accent and phrase boundary recognition [13] or emotion recognition [14]. The features are based on duration, energy, pitch, and pauses, and can be applied to locally describe arbitrary units of speech such as words or syllables. Short-time energy and fundamental frequency (F0) are computed on a frame-by-frame basis, suitably interpolated, normalized per utterance, and perceptually transformed. Their contour over the unit of analysis is represented by a handful of functionals such as maximum or slope. To account for intrinsic variation, we include normalized versions of some of the features based on energy and duration, e. g. the normalized duration of a syllable based on the average duration of the comprising phonemes and a local estimate of the speech rate. The statistics necessary for these normalization measures are estimated on all speech data of the 11 native C-AuDiT speakers (approx. 5h).

We now apply our module to different segments and construct global (utterance-level) features from that. Trying to be as exhaustive as possible, we use a highly redundant feature set (742 features) leaving it to data-driven methods to find out the relevant features and the optimal weighting of them. We compute:

**(1)** Average and standard deviation of the prosodic features derived from all *stressed syllables* (context '$0, 0$'), from all segments comprising stressed syllables and their direct successor (context '$0, +1$'), from all syllables succeeding stressed syllables (context '$+1, +1$'), and so on up to contexts '$-2, -2$' and '$+2, +2$'. The same is done for just the nuclei of stressed syllables. These features can be interpreted to generically capture isochrony properties inspired by [15].

**(2)** Average and standard deviation of the prosodic features derived from all words (context '$0, 0$'), and from all segments comprising two words (context '$0, 1$'). The same is done for syllables and nuclei. These features can be interpreted as generalizations of the deltas and proportions proposed by [5].

**(3)** Average of the absolute differences between the prosodic features derived from consecutive units. This is done for contexts '$0, 0$' and '$0, 1$' of all words, syllables and nuclei. These features can be interpreted to generalize the pairwise variability indices proposed by [4].

## 3.2. GMM-UBM Features (*gmm*)

Using the GMM-UBM approach in combination with Support Vector Machine (SVM) classification can be regarded as standard approach for speaker identification [8] and classification of speaker characteristics. We used this approach to detect the degree of pathology (or speech intelligibility/voice quality) in speakers with laryngeal cancer [16]. The system was motivated by the idea that the acoustic space of the speaker, represented by the GMM, differs from that of a healthy speaker in case of a pathology. This difference has been shown to contain information about the degree of deviation. We assume an analogue behaviour for non-native speakers and hope that the acoustic space populated by a speaker will yield contain information on the (degree of) non-native deviation from native speech.

Short-time spectral features, so-called RPLP [17], a revised variant of Hermansky's Perceptual Linear Prediction (PLP) approach [18], are used in this study. By employing the Mel filterbank instead of the Bark filter-bank and other simplifications, these features are very similar to Mel Frequency Cepstrum Coefficients (MFFCs). The sole difference is that RPLP performs

an additional spectral smoothing step that emphasizes the spectral envelope, which possibly results in a better representation of the vocal tract. This might be an explanation why we found RPLP to be a trifle ahead of MFCC in previous tasks. We use a 25 msec Hann window and a step size of 10 msec. We compute 13 cepstral coefficients and add delta coefficients (as commonly done, by regression over 5 cepstral coefficients) and acceleration coefficients (regression over 9 delta coefficients, instead of the commonly used 5), arriving at a total of 39 features. The increased context for acceleration gave us some improvements in previous tasks. We apply cepstral mean subtraction (per utterance) and discard non-speech frames using a HMM-based phoneme classifier.

A single, speaker-independent Gaussian Mixture Model (GMM) with diagonal covariances is trained on the speech data of the 11 native C-AuDiT speakers (approx. 5h), yielding the so-called Universal Background Model (UBM)[1]. This model acts as a reference model of correctly uttered speech. The model parameters are estimated in an unsupervised iterative manner by the Expectation-Maximization (EM) algorithm in 10 iteration steps. The actual speaker model is derived by adapting the parameters of the UBM to the data of the target speaker by Maximum A Posteriori adaptation. This results in a speaker-specific GMM with the parameters $\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, i = 1, ..., K$, where $K$ denotes the number of mixture components. In the basic approach a speaker is represented by a concatenation of the mean vectors $\mu_i$, the so-called GMM supervector, with dimension $39 \cdot K$. Including weights and covariances into the supervector, resulting in dimension $79 \cdot K$, might help to model the variation/distance of an L2 learner from native speakers.

### 3.3. openSMILE Features (*smile*)

OpenSMILE [10] is a toolkit for computing general-purpose acoustic and prosodic features proven successful for a variety of paralinguistic tasks. A multitude of low-level descriptors such as loudness, pitch or energy in spectral bands is modelled by many different functionals such as mean, standard deviation or quantiles. In the default configuration, which we use, no additional information such as a transcription is necessary; segments are determined automatically based on energy (or voicing, for pitch related features). Given this out-of-the-box functionality, and its success in related areas, openSMILE suggested itself as a 'must-try' candidate for our task. We employ the official feature set used in the 2011 Interspeech Speaker-State-Challenge [11] resulting in 4368 features per utterance.

## 4. Modelling and Fusion

Prediction of the continuous target scores is done by Support Vector Regression (SVR), using WEKA [19]. We use a linear kernel; the complexity parameter $C$ was optimized (up to a power of ten) for each feature set on the whole database (0.01 for pros, 1 for gmm, 0.001 for smile). System performance is estimated in leave-one-speaker-out cross-validation and reported in terms of Pearson correlation $r$. For lack of space, we omit Spearman – the values are similar except for *int* where Spearman exceeds Pearson a bit. For combining two or more of the *pros*, *gmm* and *smile* feature sets, we use late fusion, i. e. combine the outputs of separately trained SVR-Systems linearly. The weights for the combination are tuned for best performance within $\{0.0, 0.1, \ldots, 1.0\}$, again on the whole database.

[1] This is relatively few data for a UBM, but to start with and to keep things simple, we wanted to work with our available in-domain data. More native material can, at a later stage, be added easily as no annotation is needed here.

Table 1: Performance (Pearson correlation) of the GMM-UBM-System for the different target scores when employing different GMM-parameters and different numbers of mixture components $K$. The last column averages performance across all four criteria.

| Supervector | $K$ | int | acc | mel | rhy | ∅ |
|---|---|---|---|---|---|---|
| $\boldsymbol{\mu}_i$ | 64 | .520 | .566 | .522 | .567 | .544 |
| $\boldsymbol{\mu}_i$ | 128 | .503 | .583 | .559 | .549 | .548 |
| $\boldsymbol{\mu}_i$ | 256 | .495 | .582 | .551 | .552 | .545 |
| $\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ | 32 | .515 | .618 | .596 | .591 | .580 |
| $\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ | 64 | **.603** | .618 | **.620** | **.639** | **.620** |
| $\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ | 128 | .568 | **.635** | .613 | .605 | .605 |

Table 2: Performance (Pearson correlation) of the different feature sets and of their combination via late fusion.

| pros | gmm | smile | int | acc | mel | rhy |
|---|---|---|---|---|---|---|
| • | | | .590 | .595 | .735 | .793 |
| | • | | .603 | .618 | .620 | .639 |
| | | • | .436 | .428 | .524 | .538 |
| • | • | | **.683** | **.693** | **.771** | **.822** |
| • | | • | .608 | .611 | .739 | .793 |
| | • | • | .605 | .621 | .642 | .659 |

Optimizing $C$ and the weights on all data is effectively tuning on the test set; however, the effect of overfitting should be small since only 2 to 3 parameters are fitted, and only very coarsely. It proved however crucial to apply *leave-one-speaker-out* cross-validation for SVR parameter selection, as especially the *gmm* and *smile* systems showed a considerable tendency to overfit to the speakers in train when just using standard cross-validation across all instances (i. e. utterances).

## 5. Results

As this is the first time we applied the GMM-UBM-supervector approach to this data, we first run a set of experiments to find out suitable parameters. Table 1 lists the results. We got the best overall performance when using all parameters of the GMM ($\omega_i$, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$) with 64 mixture densities, cf. the penultimate row of Table 1. For the example of *rhy*, the system's output and reference are correlated with $r$=0.639.

Table 2 compares these best *gmm* results (replicated in row 3) with the performance of *pros* (row 2) and *smile* (row 4). For the more segmental scores *int* and *acc*, the *pros* and *gmm* features perform similar, e. g. for *acc*, $r$=0.595 and 0.618, respectively. For the prosody-related scores *mel* and *rhy*, however, the *pros* features are considerably better, e. g. $r$=0.793 vs. 0.639 for *rhy*. The system using *smile* features behaves similar to *pros* in so far as it can model the more supra-segmental *mel* and *rhy* scores better than *int*/*acc*, e. g. $r$=0.538 vs. 0.428 for *rhy* vs. *acc*. However, in absolute terms, performance is much lower: for *mel*, for instance, *smile* only reaches $r$=0.524 while *pros* achieves 0.735.

Rows 5–7 show the performance when combining two feature sets (cf. Section 4). All sets benefit from this 'collaboration': the fusion of two sets always yields better results than each of the sets alone.[2] For example, for *rhy*, the combination of *gmm* and *smile* (last row) yields $r$=0.659, better than the

[2] An alternative way of putting this is: when combining two feature sets, the weight 0.0 was never chosen.

stand-alone performances 0.639 and 0.538. The combination of *pros* and *gmm* is clearly best for all target scores, cf. row 5 (the results in bold face). The chosen weights for this fusion are 0.5/0.5 for *int* and *acc*, while for *mel* and *rhy*, *pros* is weighted a bit higher (0.6 vs. 0.4 for *gmm*). Depending on one's choice of $\alpha$, this fusion can be considered significantly better (one-sided test) than the best stand-alone systems:

*int*: $r$=0.683 $>$ the *gmm* system's 0.603 ($p$=0.008),
*acc*: $r$=0.693 $>$ the *gmm* system's 0.618 ($p$=0.02),
*mel*: $r$=0.771 $>$ the *pros* system's 0.735 ($p$=0.1), and
*rhy*: $r$=0.822 $>$ the *pros* system's 0.793 ($p$=0.1).

Combining all three feature sets did not yield further improvement over the combination of *pros* and *gmm*[3].

## 6. Discussion

As standalone systems, neither the *gmm* features nor the *smile* features reach a performance comparable to the prosodic feature set *pros*. This can be explained by the fact that only the latter is provided with information on the syllabic and phonetic structure of the utterances, but it can also be taken as a proof that we are on the right track with our present approach.

However, the fact that each combination of two feature sets leads to an improvement over the stand-alone performances indicates that all three feature sets contain useful, complementary information. Combining *pros* and *gmm* worked best by far; this can be explained partly by their better stand-alone performance, but it also reveals a relatively high complementarity. For all target scores, this combination yields pronouncedly better results than any of the feature sets alone. For the case of *rhy*, we could thus improve our performance from $r$=0.793 to $r$=0.822 by adding the GMM-UBM-system to our prosodic feature analysis, a 3.7% relative improvement (*mel*: 0.735 $\rightarrow$ 0.771, 4.9% rel.). Thus it can be useful for prosodic assessment to 'combine the best of two worlds' – incorporating rather 'alien' techniques such as the GMM-UBM approach which is solely based on the short-time spectral representation of speech and could therefore be expected to be independent of and irrelevant for prosody.

The fact that *smile*—i. e. openSMILE in the 'blind', out-of-the-box mode—doesn't work so well indicates that for the assessment of non-native prosody, putting in meta-information such as syllabic/phonetic structure, or position of stresses is (as might be expected) indeed vital. In future work, we will therefore combine the two approaches and hope for additional improvements through the comprehensiveness of the feature computation of openSMILE.

A clear limitation of our work so far is that we have only shown its usefulness on *known* sentences. Thus, it is perfectly applicable, e. g., for language testing, but for CAPT one usually wants to employ a module that works well for unseen material. In the future, we will therefore study the presented approaches not only in speaker-independent but also in sentence-independent evaluation setups.

## 7. Acknowledgments

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the project *C-AuDiT* under Grant 01IS07014B, by the German Ministry of Economics (*BMWi*) in the project *AUWL* under grant KF2027104ED0, and by the German Research Foundation (*DFG*) under grant EY 15/18-2.The responsibility lies with the authors. The perception experiments were conducted by Susanne Burger and Catherine Dickie. We want to thank Andreas Maier for adapting PEAKS to our task.

[3] i. e. the (vary coarsely) optimized weight for *smile* system was 0.0 when combining all three feature sets, for all of the target scores.

## 8. References

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

[2] J. Tepperman and S. Narayanan, "Better nonnative intonation scores through prosodic theory," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1813–1816.

[3] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for ted talks," in *Proc. ICASSP, Prague, Czech Republic*, 2011, pp. 5672–5675.

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

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

[6] F. Hönig, A. Batliner, and E. Nöth, "How many labellers revisited – naïves, experts and real experts," in *Proc. SLATE*, Venice, Italy, 2011, no pagination.

[7] 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, no pagination.

[8] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, pp. 308–311, 2006.

[9] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, p. 1062–1087, 2011.

[10] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, New York, NY, USA, 2010, pp. 1459–1462.

[11] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Proc. Interspeech*, 2011, pp. 3201–3204.

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

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

[14] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of tuning – prototyping for automatic classification of emotional user states," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 489–492.

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

[16] T. Bocklet, K. Riedhammer, E. Nöth, . Eysholdt, and T. Haderlein, "Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling," *Journal of Voice*, 2012, to appear.

[17] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 2997–3000.

[18] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009.