Automatic Assessment of Non-Native Prosody – Annotation, Modelling and Evaluation

Florian Hönig, Anton Batliner, and Elmar Nöth
Pattern Recognition Lab
Universität Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
Email: {hoenig,batliner}@informatik.uni-erlangen.de

***Abstract*—The first part deals with general considerations on the evaluation of both human raters and automatic systems, employed for pronunciation assessment: How can we come closer to an unbiased, realistic estimate of their reliability, given the fallibility of human annotators and the nature of machine-learning algorithms (and researchers) that adapt, and inevitably overfit to a given training set. In the second part, we will present concrete models for the assessment of the overall rhythmic quality of the learner's speech. The methods are evaluated in detail on read and semi-spontaneous English data from the German research projects C-AuDiT and AUWL.**

I. Introduction

A. Motivation

Non-native prosodic traits limit proficiency in a second language (L2) and by that, mutual understanding. 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 successfully. 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 [4], the automatic detection of erroneous word accent positions in English as L2 is addressed. Suprasegmental native traits have been, e. g. investigated recently in basic when trying to model language-specific rhythm [5], [6]. Maybe the most important general 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'. The reason is a difference in rhythm that has been noted amongst others by [7], p. 97, who speaks about syllable timed languages such as French ("the syllables [...] recur at equal intervals of time – they are *isochronous*"), and stress-timed languages such as English ("the stressed syllables [...] are isochronous"). [5] and [6] challenge this traditional terminology because in empirical studies, such an isochrony could not be observed; they claim that it is rather a more complicated constellation where especially syllables not carrying the word accent, that are weak (schwa) in 'stress-timed' languages, are produced stronger in 'syllable-timed' languages. Thus we might expect such differences to show up in L2 learners whose native language L1 does not display the native structure of L2.

To cope with these traits, L2 teachers can use explicit feedback, i. e. denote the very pronunciation error, or implicit feedback, i. e. repeat (parts of) lessons which proved to be difficult for the learner. The same strategies are available for Computer-Assisted-Pronunciation-Training (CAPT) programs. Explicit feedback should be used – but only if there is a high recall and a low false alarm rate. However, we are still far from any 'perfect' *localization* of pronunciation errors; other things being equal, a *global* assessment (of sentences, paragraphs, or whole sessions) has higher chances to correctly indicate (types of) coarse errors the learner tends to make. If any localised assessment is available, we can use this information for giving both explicit and implicit feedback, whereas a global assessment implies the sole use of implicit feedback. Rhythm is insofar special as it is often difficult to pinpoint exactly what's wrong with a given utterance that sounds unnatural, and even more difficult to convey to the learner what exactly to improve. Thus, implicit feedback to the learner's rhythm in CAPT programs, e.g. in a say-after-me or speak-with-me (shadowing) manner may be a good option for pedagogical reasons, too.

B. Machine Learning: Evaluation

The pattern recognition approach – i. e. collect annotated data, extract suitable features, and train a supervised classification or regression module using machine learning – is a universal and powerful tool in CAPT. However, great care has to be taken when estimating the accuracy of such an approach: If the collected data are not *representative* of the intended application, a strict division into *training* and *test set* has to be used during evaluation. Otherwise, the used algorithms and choices taken by the researcher may *overfit* to the data and yield optimistic estimates of the accuracy [8, p. 19]. In CAPT research (and many other fields), the data are usually *far* from being representative, as there is no running application (yet) to draw data from, and eliciting, collecting and especially annotating is expensive. Moreover, linguistic data are *per se* an open set.

Even if the manifold pitfalls of overfitting to the data[1]

[1]feature selection, accidentally tuned to the whole dataset, may stand as a popular example

are avoided, and the evaluation is technically sound in so far as all items (*instances*) used for testing have never been used for training/tuning, the estimated accuracy may still be meaningless. The reason is that just keeping training and test set disjunct w. r. t. the *instances*, and mixing everything else[2], is not enough. In fact, each partition into training and test has to be designed in such a way that the test reflects the conditions in the eventual application. For example, in the speech recognition community it is widely recognized that train and test have to be disjunct w. r. t. speakers in order to arrive at realistic estimates of the accuracy.

Depending on the application, there may be other conditions that should be different, too. When there is more than one condition that needs to be different in training and test, the evaluation gets wasteful: either time-consuming nested cross-validation schemes have to be used or only a fraction of the hard-won data can be utilized. This becomes a problem when organizing competitions such as the INTERSPEECH 2009 Emotion Challenge [9] where cross-validation is not a practicable option. A way to avoid this problem from the start would be to collect data in (at least four) partitions that are designed in such a way that they are mutually independent w. r. t. all conditions.

For CAPT, one usually wants to employ a module that works well not only for unseen speakers, but also for unseen material [10]. We will therefore include in our evaluation not only speaker-independent settings, but also a setup where train and test is disjunct w. r. t. both speakers and sentences – this will have dramatic consequences on the resulting accuracy.

C. Generative vs. Discriminative Approaches

For evaluating the accuracy of a CAPT method, we will always need a body of data from non-native speakers that includes annotated examples of both good and bad speaking performance. For modelling, however, two basic approaches can be identified:

Generative or indirect: The model only describes what is *acceptable*, and a distance measure is used to derive a score or to decide for 'correct' or 'error'. The advantage is that data collection is far easier: we can use native speech, and more importantly, when neglecting the few errors that native speakers make as well, we do not need error annotations. For example, when applying the Goodness of Pronunciation (GOP) measure [11] to identify mispronounced phonemes, we can use just transcribed native speech to build models for correctly produced phonemes, and use (an approximation of) the a posteriori probability of the target phonemes as a similarity measure.

Discriminative or direct: The model describes *both* acceptable and unacceptable pronunciations, and the pronunciation score or the decision 'correct' or 'wrong' is a direct output of the classification or regression module. This approach has the potential for optimal accuracy but data collection is much more expensive, as enough annotated non-native speech comprising both good and bad pronunciations is needed, i. e. much much more than for the evaluation of generative approaches. For the example of detecting mispronounced phonemes, this is practically infeasible in the general case due to data sparsity resulting from coarticulation effects and the different L1s of the targeted learners. For modelling frequent errors of certain target speaker groups however, it may be the method of choice, e. g. /θ/ → /s/ e. g. /T/ → /s/ for German learners of English as L2.

In practice, both approaches are often mixed to reach satisfying accuracy with feasible effort, e.g. a generative model for correct phonemes is used but a priori probabilities for mispronunciations of the target group of speakers are included.

Assuming that modelling pronunciation quality w. r. t. rhythm is less complex than modelling segmental pronunciation, we followed the discriminative approach in the present work.

D. Annotation: General Considerations

As discussed above, we need to establish reference scores for evaluating, and possibly also for training our pronunciation scoring method. Apart from the speech data that should be annotated – type, size, (balanced, stratified) sub-samples such as male/female, degree of proficiency, etc. – the main alternatives to be chosen from is a choice between experts and 'naïve' subjects for annotation and/or perceptive evaluation, and the decision on how many people to employ for the annotation task.

1) Labeller Agreement and Multiple Labellers: The variability between labellers can be traced back to at least two main factors: first, labeller-specific *traits* such as gender, dialect, sociolect, talent for assessing speech, etc., and second, speaker-specific *states* such as boredom, interest, tiredness, illness, etc. Together, all these factors can be modelled as error whose variability is higher if less subjects are employed. Following this logic, we can define the *ground truth* as the average over infinitely many labellers (for a certain group of labellers).

How many labellers to actually employ is foremost a matter of time and money – as long as some rules of thumb are followed: if there are three or more labellers, we can use majority decisions. If there are five or more labellers, we are more safe when establishing quasi-continuous judgements from ordinal ones, based on the average score of all annotators. Intuitively, around 10 is a good figure; more than 20 are employed rather rarely. In our own experience, we found that for the task of rating prosody of L2 English speech on a continuous scale, 10 labellers already yield a very good reference: A reference A^N created by averaging over the (normalised) annotations of N labellers with an average pairwise Pearson correlation of c can be expected to be correlated to the ground truth as follows [12]:

$$\mathrm{Corr}(A^N, A^\infty) = \sqrt{c/(\frac{1}{N} + \frac{N-1}{N}c)}. \quad (1)$$

[2]as is commonly done when doing cross-validation with machine learning packages

Thus, despite of a low pairwise correlation of 0.3, averaging over 10 labellers already yielded a reference with a correlation of 0.90 to the ground truth.

2) Expert vs. Naïve Labellers: Experts being able to do a detailed annotation are rare and more expensive than naïve raters; moreover, they may be biased in some way towards their own theoretical preferences. Naïve subjects are less expensive, thus more of them can be employed, and they are less biased, but care has to be taken that the task is well-defined; moreover, we cannot expect them to be as consistent and competent as the experts. Normally, less experts are employed than naïve subjects. So far, however, there are no strict guidelines for that; recently, there seems to be a trend towards low-cost (non-expert) crowdsourcing using, for example, Amazon Mechanical Turk [13]: Snow et al. conclude that for the task of affect recognition in speech, using non-expert labels for training machine learning algorithms can be as effective as using gold standard annotations from experts. Also in [14], it was shown that a large number of annotators ('Vox Populi') creates reliable annotations.

In our experiments on rating L2 German speech with respect to prosody, we compared three groups of native labellers with different expertise: naïves, phoneticians, and phoneticians with extensive labelling experience with the actual database ('real' experts) [15]. As expected, the consistency rose with the level of expertise: Regardless of whether we aim at the ground truth of experts, phoneticians or naïves, when only one labeller is employed, an expert is always the best choice and a naïve labeller the worst choice. However, when employing more labellers, good correlations to any of the three 'ground truths' can be achieved by all labeller groups. If c is d are the average pairwise correlation within two groups $\mathbb{A}$ and $\mathbb{B}$, respectively, and e the average pairwise correlation between a pair of labellers from the two groups, the correlation of N averaged labellers A^N from $\mathbb{A}$ with the ground truth B^∞ of $\mathbb{B}$ can be expected to be

$$\mathrm{Corr}(A^N, B^\infty) = \frac{e}{\sqrt{\frac{1}{N} + \frac{N-1}{N} c} \cdot \sqrt{d}}. \qquad (2)$$

Thus, we observed in our German data when employing at least five labellers of any of the three groups, very good references result that can be expected to be correlated to the ground truth of any of the three groups with at least 0.94.

3) Different L1 Dialect Backgrounds: Another aspect that may influence the perception and thus the rating of non-native speech is the background of the labellers with respect to their L1 variety. For our English data, we compared native speakers of American, British and Scottish English [16]. We observed slight differences in the perception of non-native accent, but practically no difference in scores for intelligibility or prosody. Thus we can speculate that irrespective of their own dialect or regional accent, annotators have internalised a common standard of their own L1.

4) Correlated Scores: When collecting labels for different scales such as intelligibility and prosody, one will usually find that the ratings are correlated among themselves to some extent. In this context, it is desirable to abstract from individual labeller variability. We can do this by estimating the correlation between the ground truths of the scales: if c and d are the average pairwise correlations within the labels for each of the two scales, and e the average correlation between one labeller's first scale with another labeller's second scale, we can use Equation 2 with $N \to \infty$, i. e. $e/\sqrt{cd}$.

5) Weighting Labellers: Even within a homogeneous group of labellers there will be individual differences regarding talent, diligence, time spent on task etc. which will have an impact on the quality of the annotations. Therefore an obvious possibility to improve the quality of the reference is to assign weights when averaging multiple annotations. A basic and robust approach is choosing the correlation to the other labellers as a weight [17]. In [18], a maximum likelihood estimator is derived that estimates weights and the combined reference in an iterative manner. In our own experiments, we use a similar but simplified, and more stable approach: using initially uniform weights, we estimate the mean square error of the labeller w. r. t. the ground truth, and set the weight indirectly proportional. On our German and English prosody scores, however, we did not see a big improvement by using weights, with neither of the three mentioned methods. For example on our new semi-spontaneous dialogue data (see below), we could improve the expected correlation to the ground truth just a little bit from 0.85 to 0.86. A reason for this may be that we hired our labellers in the traditional way with personal contacts, and paid them well, so we did not encounter problems that are reported for using e. g. Amazon Mechanical Turk where one has to check for 'spammers' who try to get the money without doing any real annotation.

6) Paying Labellers: When planning the annotation of a database, it is very convenient to pay piece-work (i. e. per annotated time, and not per annotation time). Often the money available for annotation is fixed, and then one can calculate what portion of the data one can afford to have annotated with how many labellers. To be ethically acceptable, the payment should not be too low; still, the quality of the annotation may suffer because some labellers may try to finish the job as quickly as possible. Other labellers take more time; paying them the same is unfair twice, as the slower labellers tend to deliver higher quality. Another possibility would be to pay per quality, where a quality measure could be derived in the same way as the weights when creating a combined reference by weighted averaging. In the annotation during the AUWL project, we observed a correlation of 0.87 between weights (assigned for intelligibility, non-native accent and prosody) and the time spent by the labellers. Although we cannot call this significant (only five labellers for this task), this is a strong trend. Nevertheless, it is still questionable to use that as a basis for payment: if the quality measure is normalized with respect to all labellers, labellers will effectively compete against each other, and if the quality measure is absolute, payment will be less when the task is difficult. Summing up, the best way still seems to pay an hourly wage, if organisational constraints allow for that.

In the annotation for the AUWL project, we could only pay piecework. However, after completion, we realized that one labeller – the best – took

almost twice as much time as the rest, so we decided to pay an extra compensation.

7) Comparing Humans and Machines: For judging the estimated accuracy of an automatic system, it is instructive to compare with human performance. In order to do that in a fair manner, one should be aware that the correlation of the system with the reference is at best an optimistic estimate of accuracy. In the end, we want to know how well a system predicts a certain (abstract) *score*, not the collected imperfect *reference*, so we can for example not claim that any system performs better than the combined labellers. In fact, the final accuracy of a system should be estimated as the correlation between system output Y and ground truth A^∞:

$$\mathrm{Corr}(Y, A^\infty) \quad = \quad \mathrm{Corr}(Y, A^N) \cdot \mathrm{Corr}(A^N, A^\infty). \quad (3)$$

Consider for example a hypothetical system that is trained with the average of five labellers with a pairwise correlation of 0.5. Using Equation 1, this yields a reference with an expected correlation to the ground truth of ≈ 0.91. At the first glance, an automatic system that correlates with the reference with 0.6 seems better than the average human. However, as argued above, the correlation of the system with the ground truth can only be expected to be $0.6 \cdot 0.91 \approx 0.55$. On the other hand, a single labeller can be expected to correlate with the ground truth with $\sqrt{0.5} \approx 0.71$. Thus, we should be careful not to underestimate human performance or overestimate the performance of our systems.

II. Databases

For the present work, we use two databases: Read English material from our German research project C-AuDiT and the EU project ISLE [19], and new semi-spontaneous English data from research project AUWL, collected with the help of our dialogue training tool *Dialogue of the Day* (dod).

A. C-AuDiT

Read material is, of course, less naturalistic than spontaneous one; however, it has two advantages: First, it is easier to process, and second, it allows incorporation into existing automatic training software which still builds upon written and read data. Thus, it is a relevant object of study, also from the point of view of an commercial applicant of CAPT.

1) Material and Speakers: We recorded 58 English L2 speakers: 26 German, 10 French, 10 Spanish, 10 Italian and two Hindi speakers, and additionally 11 native American English (AE) 'reference' speakers. They had to read aloud 329 utterances shown on the screen display of an automated recording software, and were allowed to repeat their production in case of false starts etc. The data to be recorded consisted of two short stories (broken down into sentences to be displayed on the screen), sentences containing, amongst other, different types of phenomena such as intonation or position of phrase accent (*This is a house.* vs. *Is this really a house?*), or tongue-twisters, and words/phrases such as *Arabic/Arabia/The Arab World/In Saudi-Arabia, ...*; pairs such as *'subject* vs. *sub'ject* had to be repeated after the prerecorded production of a tutor. Where applicable, an expert annotated a likely distribution of primary and secondary phrase accents and B2/B3 boundaries [20] of a prototypical, articulate realisation.

When designing our recordings, we took 30 sentences from the ISLE database [19], which contains non-native English from 26 German and 26 Italian speakers. From this intersection, we defined the subset of the following five sentences that were judged as 'prosodically most error-prone for L2 speakers of English' by three experienced labellers [4]:

> *We're planning to travel to Egypt for a week or so.*
> *Can I have soup, then lamb with boiled potatoes, green beans and a glass of red wine?*
> *They will have to transport the components overland.*
> *The referee needed a police escort after the match.*
> *The company expects to increase its workforce next year.*

2) Annotation: Taking all speakers from C-AuDiT and ISLE that spoke all five sentences, we arrived at approx. one hour of speech from 94 speakers. Using the tool PEAKS [21], the annotation was conducted as a web-based perception experiment. Twenty-two native AE, 19 native British English (BE), and 21 native Scottish English (SE) speakers with normal hearing abilities judged each sentence in random order regarding different criteria, answering questions on intelligibility, non-native accent and the following two questions on prosody on a five-point Likert-scale:

- THIS SENTENCE'S MELODY SOUNDS...
  (1) *normal* (2) *acceptable, but not perfectly normal* (3) *slightly unusual* (4) *unusual* (5) *very unusual*
- THE ENGLISH LANGUAGE HAS A CHARACTERISTIC RHYTHM (TIMING OF THE SYLLABLES). HOW DO YOU ASSESS THE RHYTHM OF THIS SENTENCE?
  (1) *normal* (2) *acceptable, but not perfectly normal* (3) *slightly unusual* (4) *unusual* (5) *very unusual*

As already mentioned above, we found no real difference between the ratings from the AE, BE, or SE labeller, so we lumped them all together to get a single combined score for each utterance. It turned out that these combined ratings for *mel* and *rhy* are highly correlated among themselves with 0.95. So the question was whether the labeller were at all able to distinguish between the two. To answer this, we estimated the correlation between the ground truth of the scores as described in Section I-D4. Thus, we can expect the ground truths of *mel* and *rhy* to correlate with 0.97. Interestingly, we found the British labellers to behave a bit different (0.95; AE: 0.98, SE: 0.99). Our conclusion is that there may be a small difference, but too small to be considered for automatic assessment at the current state of the art. Thus, we decided for the present study to form a combined rating *pros* by averaging the 124 (normalized) annotations of both the *mel* and *rhy* scores. The expected correlation of this combined score with its ground truth is 0.99.

B. Dialogue of the Day (dod)

Reading leads to a special speaking style and can have a disruptive effect on speech, especially for learners with low L2 competence. Therefore, we took a different approach to data collection in our research project AUWL and designed a client-server tool for practising pre-scripted dialogues.

1) Training Tool: Before embarking on the dialogue training, the learner can first listen to the whole dialogue spoken by reference speakers. Then the learner enacts the dialogue with a reference speaker as a dialogue partner. In doing so, he can either have his lines prompted by a reference speaker and repeat afterwards, or directly read them off the screen (karaoke), or speak simultaneously with a reference speaker (shadowing). For facilitating shadowing, it can optionally be combined with prompting. Taking into account less proficient learners, one can choose between reference recordings spoken in a normal or in a slow tempo, and longer dialog steps can be subdivided. Options for choosing from different reference speakers, swapping roles, (re-)starting from an arbitrary position, replaying the latest own version of a dialog step, replaying the whole enacted dialog, or using own recordings for the dialog partner, complete the versatile training tool which is admittedly too complicated for end customers. Using this tool, we were able to elicit application-relevant speech which is considerably more spontaneous and less reading-style.

2) Material: We created 18 dialogues on topics such as business negotiations, shopping or holidays, with six for each of the CEF [22] levels A2 (elementary), B1 (pre-intermediate), und B2 (intermediate). Three female and three male professional native speakers spoke the material in both normal and slow tempo, resulting in 1908 recorded reference utterances or 2.2 hours of speech. As for the C-AuDiT material, we annotated a likely distribution of primary and secondary phrase accents and B2/B3 boundaries of a prototypical, articulate realisation for each dialogue. Possible points for subdivisions were annotated independently, because the presumptive B3 boundaries annotated were not suitable in some cases. The audio tracks to be replayed for the subdivided mode were created by automatically cutting the whole recordings, using a speech recognizer for segmentation.

3) Speakers: We started with 85 volunteering learners, who got a login for a web-based system where they could use the training tool in a self self-dependent manner. The learners were free in the choice of the dialogs and training modes such as shadowing. All recordings and dialogue timing information were stored at the server. Although we asked the learners to use a headset, the resulting audio quality was quite heterogeneous. At the end of the data collection, we got usable speech material from 31 speakers. According to a self-assessment, CEF levels are distributed as follows: 5×A2, 5×B1, 10×B2, and 11×C1. In total, the material amounts to 5145 utterances in 1019 dialog runs or 7.8 hours of speech. Each utterance was classified by a single annotator into 'clean' (5.5h), 'usable' (mainly usable content, but word editing and louder noise such as coughing; 1.7h), or 'unusable' (unusable content or dominant noise due to wrong audio settings etc.; 0.6h).

4) Annotation: The clean and usable material was annotated by five native post-graduate phoneticians. As with the C-AuDiT material, we asked questions on intelligibility and non-native accent on a five-point Likert scale, but according to our experience with the *mel* and *rhy* scores, we just asked one merged question regarding prosody (*pros*):

> THE ENGLISH LANGUAGE HAS A CHARACTERISTIC PROSODY (SENTENCE MELODY, AND RHYTHM, I.E. TIMING OF THE SYLLABLES). THIS SENTENCE'S PROSODY SOUNDS…
> (1) *normal* (2) *acceptable, but not perfectly normal*
> (3) *slightly unusual* (4) *unusual* (5) *very unusual*

We normalized and averaged the five annotations to get a single score for each utterance; for *pros* its expected correlation to the ground truth is 0.85. Additionally, the labellers had to mark words or parts of a sentence with a particularly unusual/non-native prosody. We measured the time the labellers took for annotation: we observed real-time factors between 2.2 and 7.5 (average: 3.9). For the present evaluations we only use the utterances classified as clean.

III. Prosodic Features

In order to obtain suitable input parameters for an automatic prosody assessment system, we compute a prosodic 'fingerprint' of each utterance. All processing is done fully automatic; however, we assume that the spoken word sequence is identical with the utterance the speaker had to read. 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 [16] and adapted to utterance level instead of speaker level.

A. Specialized Rhythm Features

There is a body of research on modelling language-specific (native) rhythm. These hand-crafted, specialized parameters are promising candidates for our task.

1) Duration Features (Dur): A basic but fundamental property of speech is how fast something is said. We compute the average duration of all syllables of the utterance, and the average duration of vocalic and consonantal intervals (two features).

2) Isochrony Features (Iso): In order to capture possible isochrony properties [7], we calculate distances between centres of consecutive stressed or consecutive unstressed syllables. The centres are identified as the frames with maximal short-time energy within a nucleus. We compute six features: mean distances between stressed, and between unstressed syllables, standard deviations of those distances, and the ratios of those means and standard deviations.

3) Variability Indices (PVI): Following [5], we identify vocalic and consonantal intervals and calculate the raw Pairwise Variability Index (rPVI) which is defined as the absolute difference in duration of consecutive segments and its normalised

version nPVI (rPVI divided by the mean duration of the segments) for vocalic and consonantal segments. Additionally, following [23], we compute the control/compensation index (CCI) for vocalic and consonantal segments. This variant of rPVI takes into account the number of segments[3] composing the intervals. In total, six PVI features are computed.

4) Global Interval Proportions (GPI): Following [6], we compute the percentage of vocalic intervals (of the total duration of vocalic and consonantal intervals), and the 'Deltas': standard deviation of the duration of vocalic and consonantal intervals. Additionally, we include variation coefficients ('Varco') [24] for vocalic and consonantal intervals, i. e. normalized versions of the deltas. Together, we compute five Global Proportions of Intervals.

5) Combination of All Rhythm Features: Later in the experimental evaluation, these feature groups will either be analysed individually, or pooled (*Rhy-All*, 19 features).

B. General-Purpose Prosodic Features (Pros)

The expert-driven, specialized features described above are all based on duration, so they might miss other relevant information present in the speech data, such as pitch or loudness.

Therefore, we tried to capture as much potentially relevant prosodic information of an utterance as possible in an approach somewhere between knowledge-based and brute-force.

1) Local Features: We first apply our comprehensive general-purpose prosody module [25] which has proven suitable for various tasks such as phrase accent and phrase boundary recognition [25] or emotion recognition [26]. 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 the native data of each corpus (11 native speakers amounting to five hours for *C-AuDiT*; 6 native speakers in two different tempi amounting to 2.2 hours for dod).

2) Global Features: We now apply our module to different units 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:

- 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 [7].
- 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 [6], [24].
- 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 [5], [23].

C. Combination of all Global Features

The combination of *Rhy-All* and *Pros* will be referred to as *All* in the evaluation.

IV. Modelling

The collected Likert scores for prosody are discrete random variables with five possible values. One option would therefore be to formulate the automatic assessment task as a five-class classification problem. However, we chose to automatically assess the pronunciation on a continuous scale, i. e. regression for the two reasons:

- When merging multiple labellers to get a reliable reference, information is lost, the more so as the ratings 'unusual'/'very unusual' are chosen rarely. Averaging the scores to get a quasi-continuous reference solves this problem.
- Classification does not reflect the ordinal nature of the labels.

A. Global Model

Our first approach was to take a number of utterance-level features as described in Section III and feed them, together with the reference values, to a suitable machine learning algorithm for regression. We chose Support Vector Regression (SVR), using WEKA [27]. With a suitably chosen complexity parameter C and kernel function, one can achieve both linear and non-linear models with good generalization ability even in the presence of many features.

B. Local Model

Our alternative approach uses a divide-and-conquer strategy: First, all syllables of an utterance are scored individually; the resulting scores are then combined by averaging[4]. For predicting a syllables score, we again apply SVR. Because we do not have a syllable level annotation, we use the score for

[3] Usually a phoneme is one segment; exceptions are e. g. long vowels which count as two segments.

[4] Our efforts to do a more intelligent, weighted fusion by estimating confidences were unsuccessful so far.

the whole utterance as a target for each syllable. The following features are used:

- The general purpose prosodic features as described in Section III-B1 for the current syllable, for its nucleus, and for the word the syllable belongs to, contexts '$-2,-2$', '$-2,-1$', ..., '$+2,+2$' (312 features),
- mostly binary features encoding primary/secondary word accent and prototypical phrase boundaries in the neighbourhood of ± 2 syllables, position of the current syllable witin the word and utterance (60 features),
- features encoding prototypical phrase accent, number of syllables, position in utterance etc. of the word the current syllable belongs to (10 features), and
- the number of words, and the sentence mood (statement, exclamation, question) of the utterance (four features).

Again, we tried to capture all potentially relevant information, accepting high redundancy within the feature set, and leaving it to machine learning algorithms to find out the actually relevant ones. We hope that this new divide-and-conquer may:

- Possibly provide a higher robustness because the SVR is trained with more instances and less features than in global model, and the final utterance score is an average over many single scores;
- capture information that is lost when levelling down the utterance to the global features as described in Section III-B2, by the precise, chronological context of a syllable represented by the features, and

A weak spot is definitely the use of utterance level scores. We tried to obtain syllable scores by a bootstrapping approach; we got promising but no conclusive results yet.

An obvious extension of the approach is to enrich each syllable's features with the global utterance features *Pros* or *All*. Accordingly, the local models will be referenced as *Local*, *Local+Pros* and *Local+All* in the evaluation.

V. Experiments and Results

A. Nuisance Removal

As discussed, obtaining representative training/testing data is difficult. Sometimes, however, there may be prior knowledge that might help to make the data more representative, or the evaluation more realistically, e. g. a known invariance that can be taken into account. When rating rhythm, such an invariance might be the speaking rate: Ideally, pronunciation scores should be invariant against tempo (within reasonable limits); after all, native speakers can speak fast or slow and should get good scores always. On the other hand, good learners (with good pronunciation scores) tend to speak faster than poor learners (with worse pronunciation scores). Thus, tempo is definitely a useful feature for automatic scoring. So when building a system for application, one will want to utilize tempo, but for judging the aptness of features, it can be interesting to study what happens when ignoring it. In order to do so, we estimate tempo by the average syllable duration T, and alter the *reference* Y such that it is no more correlated to T:

$$Y' = Y - \frac{\mathrm{Cov}(Y,T)}{\mathrm{Var}(T)} \cdot T. \tag{4}$$

For the C-AuDiT data, the syllable duration correlated indeed strongly (0.59) with the *pros* rating, which might partly be explained by the reading style or reading-related difficulties of the learners. Removing the correlation to duration from the labels did not affect the reliability much: the expected correlation of the combined labels to the ground truth fell from 0.99 to 0.98. For dod, duration is not correlated so strongly with *pros* (0.23), but the expected correlation of the combined labels to the ground truth suffered a bit because of fewer available labellers (five): it dropped from 0.85 to 0.80.

B. Choice of Meta-Parameters

In order to get useful results with our model, it is vital to choose suitable meta-parameters for SVR. All input features are transformed to lie within $[0;1]$ as commonly done. As kernel functions, we considered only the linear kernel and the normalized polynomial kernel [28] with lower orders, i. e.

$$\tilde{\mathrm{K}}(\boldsymbol{x},\boldsymbol{y}) = \frac{\mathrm{K}(\boldsymbol{x},\boldsymbol{y})}{\sqrt{\mathrm{K}(\boldsymbol{x},\boldsymbol{x})\mathrm{K}(\boldsymbol{y},\boldsymbol{y})}} \tag{5}$$

$$\text{with} \quad \mathrm{K}(\boldsymbol{x},\boldsymbol{y}) = (\langle \boldsymbol{x},\boldsymbol{y}\rangle + 1)^p \tag{6}$$

for exponent $p \in \{4, 8\}$. In our experience, this normalization of the vectors in feature space tends to improve performance, shorten training times and facilitate the optimization of the complexity parameter C. We try all combinations of $C \in \{0.001, 0.01, 0.1, 1\}$ and the three kernels and only report the best result. Strictly speaking, we would have to optimize these meta-parameters automatically inside a cross-validation loop, but given the need to evaluate speaker- and utterance-independently (see below), this would entail a 4-fold *nested* cross-validation, i. e. prohibitive computational costs. Thus, we effectively optimize the meta-parameters on the test set, but the effect of overfitting should be small since we are only optimizing two parameters, and only very coarsely.

C. Speaker-Independent Evaluation

In a first set of experiments, we calculate the accuracy of different systems in a two-fold *speaker-independent* cross-validation, i. e. in each of the two folds, half of the data is used for training, and the other half – disjunct w. r. t. speakers – is used for testing.

D. Speaker- and Utterance-Independent Evaluation

In a second set of experiments, we evaluate *speaker- and utterance-independently*, i. e. each pair of training and test set has to be disjunct with respect to speakers *and* utterances. For C-AuDiT, we perform a five-fold utterance (i. e. leave-one-out) cross-validation within a two-fold speaker cross-validation, i. e. for each of the $2 \times 5 = 10$ folds, $4/5 \times 1/2 = 40\%$ of the data can be used for training, $1/5 \times 1/2 = 10\%$ for testing, and 50% cannot be used. For dod, we perform just a two-fold utterance cross-validation within a two-fold speaker cross-validation, i. e. for each of the $2 \times 2 = 4$ folds, $1/2 \times 1/2 = 25\%$

of the data can be used for training, the same amount for testing, and 50% cannot be used.

E. Results

Results for different feature sets and models are given in Table I.

1) C-AuDiT: The first row (not counting headlines) of Table I refers to the most optimistic evaluation criteria: unchanged reference scores, and just speaker-independent evaluation, i. e. estimated performance for an assessment task on previously known sentences. Here, all feature sets and models, apart from *GPI* already yield relatively high correlations $\geq$ 0.58. *Rhy-All* (0.73) already scores almost the maximally reached correlation (0.75: *Pros*, *All*, *Local+Pros*, *Local+All*).

We now study how the feature sets perform when the plainest feature, the average syllable duration, is excluded from competition by making the reference uncorrelated to it, see the second row in Table I. For the *Iso* features (0.22) we now see that the previous success (0.58) was largely owed to also coding duration. The other rhythm features *GPI* (0.34) and especially *PVI* (0.45) are more successful in coding rhythm quality beyond tempo. The combination, *Rhy-All* yields no less than 0.54. The 'brute-force' approach *Pros* now is a bit clearer ahead of its 'expert-tailored' competitor *Rhy-All* with 0.58, and *Local+Pros* and *Local+All* are in the lead by a whisker with 0.59.

When evaluating also sentence-independently, results generally drop quite dramatically (see third and fourth row of Table I). When allowing the use of duration (third row), i. e. estimate the performance for an assessment task on arbitrary sentences, *Rhy-All* now only scores 0.58 (vs. 0.73 in sentence-dependent evaluation), and *Pros* and *All* drop even further to 0.53 (vs. 0.75). Apparently, the high number of features compared to the number of training instances ($50\% \times 5 * 94 = 235$) presents a problem here for generalizing to unknown sentences, otherwise *All* (which includes *Rhy-All*) would not score worse (0.53) than *Rhy-All* and *Local+Pros* and *Local+All* (0.54) would not be worse than *Local* (0.64). This bring us to the best performing approach in this setting, *Local*, which scores still 0.64 in this setting. Obviously, here the efforts for more robustness IV-B bear fruit. Combining *Pros* and *Local* by late fusion (not contained in Table I; weights – very coarsely – optimized on test: 0.3 resp. 0.7), we can further improve the correlation to 0.67.

Looking at the modelling power beyond duration (fourth row), *Rhy-All* is almost at the top (0.33), clearly beating *Pros* (0.26), presumably again caused by too few training instances to take advantage of the feature set. At least, combining the local with the global approach (*Local+Pros*) catches up (0.33 too), and *Local+All* just manages to be in the lead with 0.34.

2) dod: Here, results show a similar pattern, but sentence-dependent performance (fifth and sixth row) is only a bit better than sentence-independent performance (seventh and last row) is much less pronounced, which is due to the fact that *dod* contains much more different sentences (410 vs. 5). This reduces the danger of overfitting to the sentences (or the ability to adapt, for text-dependent tasks). Also, the difference between original (rows five and seven) and duration-deprived reference (sixth and last rows) is less clear, since we have seen that duration is only correlated to *pros* with 0.23 on *dod*.

The relevant results for a sentence-dependent assessment task (row five) show just a very slight preference for *Pros* (0.57) when compared to *Rhy-All* (0.56). Also for a sentence-independent assessment task (row seven), *Pros* (0.52) is only a little ahead of *Rhy-All* (0.49). The divide-and-conquer approach was less successful: *Local* did not score more than 0.48 here. Late fusion of *Pros* with *Local* improved the result by a fraction to 0.53 (not contained in Table I; weights: 0.3 resp. 0.7).

Regarding modelling power beyond duration, *Pros* (0.47) could again be shown to be noticeably better than *Rhy-All* (0.42). As for *C-AuDiT*, *Local+All* just manages to be in the lead with 0.48. For this most difficult task, the best results are clearly ahead of those of *C-AuDiT* (maximally 0.34, see *Local+All* in row four), which is owed to the better representativeness of *dod*.

F. Discussion

1) Language Testing: The sentence-dependent results (up to 0.75) for C-AuDiT are quite good; however, the applicability for CAPT is limited. For language *testing*, however, where only a finite set of test items is needed, it is perfectly feasible only to use sentences already contained in the training set of an automatic scoring method. Nevertheless, it is questionable whether the discriminative approach pursued is best suited for this task. After all, the needed data collection is very costly, and some adaptations (calibration/partitioning of the material) had to be employed because the method cannot adapt to too many sentences at once (see the performance drop for *dod*). However, as test items can be defined *a priori*, a generative approach based on native templates along the lines of [29] is much cheaper and may yield similar or even superior results: the number of needed test items is – compared to a CAPT application – much smaller, so one can afford to record template utterances by many speakers, presumably leading to meaningful distance measures.

2) CAPT: The best sentence-independent result for *C-AuDiT* was a correlation of 0.67 to the reference. Taking into account the quality of the reference using Equation 3, this means that the system has an expected correlation of $0.67 \cdot 0.99 = 0.66$ to the ground truth of *pros*. This is clearly better than the performance of the average individual labeller (0.54) probably owed to the fact that *C-AuDiT* is a relatively easy task due to the reading prosody/reading difficulties, which allows the automatic system, naturally adapting to the given domain, to take a 'shortcut' for rating prosody. For *dod*, the corresponding best automatic result was 0.53. Given the relatively low quality of the reference, this means that the system has a correlation of $0.53 \cdot 0.85 = 0.45$ to the ground truth. The average labeller on the other hand is clearly ahead with 0.58. Our interpretation is that *dod* is the more difficult task because reading difficulties play a smaller role, and the

TABLE I

RESULTS FOR DIFFERENT FEATURE SETS AND MODELS FOR C-AUDIT (UPPER HALF) AND DOD (LOWER HALF) IN TERMS OF PEARSON CORRELATION COEFFICENT BETWEEN THE SYSTEM'S OUTPUT AND THE REFERENCE. 'SPEAKER' STANDS FOR A SPEAKER-INDEPENDENT EVALUATION, 'SPEAKER+SENTENCE' FOR A SPEAKER- AND SENTENCE-INDEPENDENT EVALUATION. 'ORIG' REFERS TO TAKING THE ORIGINAL COMBINED RATINGS OF ALL LABELLERS AS A REFERENCE; FOR 'W/O DUR' THE CORRELATION TO THE AVERAGE SYLLABLE DURATION HAS BEEN REMOVED.

Corpus	Evaluation	Reference	*Dur*	*Iso*	*PVI*	*GPI*	*Rhy-All*	*Pros*	*All*	*Local*	*Local+Pros*	*Local+All*
C-AuDiT	Speaker	Orig	0.60	0.58	0.59	0.45	0.73	**0.75**	**0.75**	0.69	**0.75**	**0.75**
		w/o Dur	0.14	0.22	0.45	0.34	0.54	0.58	0.58	0.50	**0.59**	**0.59**
	Speaker+Sentence	Orig	0.54	0.50	0.42	0.19	0.58	0.53	0.53	**0.64**	0.54	0.54
		w/o Dur	-0.13	-0.15	0.25	0.16	0.33	0.26	0.28	0.21	0.33	**0.34**
dod	Speaker	Orig	0.48	0.50	0.41	0.37	0.56	**0.57**	**0.57**	0.53	**0.57**	**0.57**
		w/o Dur	0.40	0.43	0.38	0.34	0.50	0.52	0.52	0.50	**0.53**	**0.53**
	Speaker+Sentence	Orig	0.40	0.45	0.33	0.31	0.49	**0.52**	**0.52**	0.48	0.51	**0.52**
		w/o Dur	0.32	0.37	0.31	0.28	0.42	0.47	0.47	0.44	0.47	**0.48**

speech sounds more spontaneous. Thus, with the 'crutches' no longer available, the automatic systems are revealed to still perform with sub-human performance. However, we take some comfort in the conviction that as soon as we hire some more labellers, we will get somewhere near the performance of the average human: the quality of the reference will be increased by more labellers, and it is also likely that the system will show a higher correlation to better labels. Thus, we can expect to increase both factors of Equation 3.

VI. OUTLOOK

A. More Features

Promising approaches to feature extraction for the discriminative approach are, e. g., the GMM-UBM super-vector approaches [30] and prosodic contour features [31] developed in the field of speaker identification, or the combination of both approaches [32], [33].

B. Generative Approach

It remains to be answered whether the complexity of rhythm is not too high for the chosen discriminative approach, given the efforts required for collecting suitable data. Possibilities for generative approaches would be a counterpart to the GOP algorithm for discrete prosodic events such as boundaries and accents which can be recognized or decoded with reasonable accuracy [25], or using conditional densities to make do without discrete prosodic classes.

C. Feedback

Up to now we have concentrated only on assessment and have completely ignored feedback. It would be interesting to study whether approaches based on machine learning can contribute to giving useful feedback. An example could be assessment modules that only use specific prosodic aspects such as loudness or duration to derive more specific feedback on what the learner should concentrate on in order to improve. Another improvement would be more localized feedback; possibly, the *Local* approach could be extended by bootstrapping to derive and predict syllable-level scores.

VII. CONCLUSION

The impact of suboptimal non-native prosody on understanding is well-known and has received some attention lately. In this article, we wanted to contribute to some of the most basic questions related to this topic; to this aim, we collected and annotated two databases with English as L2, spoken by speakers of different L1. The data were (1) read or prompted. We implemented (2) specialized rhythm features suggested in the phonetic literature as well as (3) a large feature set comprising general-purpose prosodic features. We addressed (4) the differences in employing different types of more or less expert labellers and (5) different numbers of labellers, and computed, based on comparing the labeller, (6) estimates of the effective quality of averaged annotations and automatic scores. Evaluation was done (7) speaker-independently and (8) utterance-independently. We showed the relevance of steps (1) to (8), based on correlations obtained for regression models with the human reference. Eventually, we discussed the impact of the single steps on performance and usability in real-life CAPT application.

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, and by the German Ministry of Economics (*BMWi*) in the project *AUWL* under grant KF2027104ED0. The responsibility lies with the authors. The perception experiments were conducted/supervised by Susanne Burger (Pittsburgh), Catherine Dickie and Christina Schmidt (Edinburgh), and Tanja Ellbogen and Susanne Waltl (Munich). We want to thank Andreas Maier for adapting PEAKS to our task.

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] 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 *Proceedings of SLATE*, Wroxall Abbey, 2009, no pagination.

[5] 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.

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

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

[8] H. Niemann, *Klassifikation von Mustern, 2. Auflage*. Heidelberg: Springer, 2003.

[9] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech*, Brighton, 2009, pp. 312–315.

[10] H. Günther, "Zur methodischen und theoretischen Notwendigkeit zweifacher statistischer Analyse sprachpsychologischer Experimente. Mit einer Anmerkung von R. Kluwe. / methodological and theoretical arguments for two-fold statistical analysis in psycholinguistic experiments. with comments by R. Kluwe," *Sprache & Kognition*, vol. 3, no. 4, pp. 279–285, 1983.

[11] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, Univ. of Cambridge, 1999.

[12] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "How many labellers? Modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Proc. SLATE*, Tokyo, Japan, 2010, no pagination.

[13] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks," in *Proc. of the Conference on EMNLP*, Honolulu, Hawaii, 2008, pp. 254–263.

[14] W.-H. Lin and A. Hauptmann, "Vox populi annotation: Measuring intensity of ideological perspectives by aggregating group judgments," in *Proc. LREC*, Marrakesh, 2008.

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

[16] 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.

[17] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. ASRU*, San Juan, Puerto Rico, 2005, pp. 381–385.

[18] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[19] 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.

[20] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, pp. 193–222, 1998.

[21] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS - a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, pp. 425–437, 2009.

[22] Council of Europe, Ed., *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001, available as PDF from www.coe.int/portfolio, last visited 11th April 2012.

[23] P. M. Bertinetto and C. Bertini, "On modeling the rhythm of natural languages," in *Speech Prosody 2008, May 6-9, 2008, Campinas, Brazil*, 2008.

[24] V. Dellwo, "Influences of speech rate on the acoustic correlates of speech rhythm. an experimental phonetic study based on acoustic and perceptual evidence," Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, 2010.

[25] 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.

[26] 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.

[27] 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.

[28] A. B. A. Graf, A. J. Smola, and S. Borer, "Classification in a normalized feature space using support vector machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 597–605, 2003.

[29] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental english through parroting," in *Proc. Speech Prosody*, Chicago, 2010, no pagination.

[30] 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.

[31] M. Kockmann, L. Burget, and J. Černocký, "Investigations into prosodic syllable contour features for speaker recognition," in *Proc. ICASSP*. IEEE Signal Processing Society, 2010, pp. 4418–4421.

[32] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[33] M. Kockmann, "Subspace modeling of prosodic features for speaker verification," Ph.D. dissertation, Brno University of Technology, Faculty of Information Technology, 2012, to appear.