First published in the *ISCA Workshop on Speech and Language Technology in Education (SLaTE)* 2009, Wroxall, Warwickshire, U.K.

Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners

Florian Hönig[1], *Anton Batliner*[1], *Karl Weilhammer*[2], *Elmar Nöth*[1]

1 - Chair of Pattern Recognition, Department of Computer Science, Friedrich-Alexander-University Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, Germany
2 - digital publishing, München, Germany

Abstract

So far, applied research aiming at computer-assisted pronunciation training has normally concentrated on segmental aspects. Here, we present a database with realizations of non-native English speakers with German, French, Spanish, and Italian as native language. We concentrate on the acoustic-prosodic modelling of word accent position and use a large prosodic feature vector to automatically recognize erroneous word accent positions produced by non-native English speakers.

1. Introduction

The automatic pronunciation assessment of second language (L2) learners nowadays concentrates on segmental errors and their subsequent treatment in computer-assisted language learning (CALL), esp. in computer-assisted pronunciation training (CAPT) [1, 2]. However, it is not only segmental errors but also suprasegmental 'peculiarities' that impede the understanding of L2 learner's productions. Such suprasegmental native traits have been, e. g., investigated recently in basic research when trying to model language-specific traits such as rhythm [3, 4]; a few studies deal with non-native accent identification using prosodic parameters [5, 6].

Whereas rhythm and intonation are global phenomena, accentuation, i. e. word accent (stress) and phrase accent position, is rather local and confined to the respective unit, i. e. one syllable in the case of **word accents (WA)**, and (one syllable in) a word in the case of phrase accents, in relation to its immediate surrounding, i.e. the word in the case of word accents, and the phrase in the case of phrase accents. Although fine-grained graduations have been proposed, normally only two or three grades are assumed: accent vs. no accent, or primary, secondary, and no accent.

The prosodic features that are most relevant in American English (AE) for the marking of such accentuation have already been dealt with in the classic studies [7, 8]; it is duration, pitch, and energy. (Note this is just the most important prosodic features; of course, other features such as vowel quality or pause structure can be relevant as well.) The ranking of importance has been a matter of some debate but nowadays, it seems fair to conclude that energy and duration might be most important, but of course, pitch contributes, albeit to a lesser extent [9, 10].

Speakers of English as L2 can be more or less fluent, having more or less pronounced L1 traits (i.e. a foreign 'accent' in the other meaning of this word); if they have reached some degree of fluency, they often run the risk of not being corrected as for their English pronunciation; that way, erroneous WA placement can survive, having a strong impact on the audience. Such 'false friends', i.e. erroneous L1-L2 transfer of **word accent position (WAP)**, can be found e.g. for *'category* vs. French *ca'tégorie*, or for *an'alysis* vs. German *Ana'lyse*.[1] It will make sense to try and avoid such **islands of failure** which tend to get fossilized, at a rather early stage of learning.

In this paper, we want to present first results from the German research project C-AuDiT (Computer-Aided Pronunciation and Dialogue Training) aiming at employing prosodic features for the recognition of wrong WAP.

2. Material and annotation

We recorded 56 English L2 speakers: 25 German, 11 French, 10 Spanish, and 10 Italian speakers, and additionally four native AE 'reference' speakers. They had to read aloud 329 utterances shown on the screen display of an automated recording software; they were allowed to repeat their production in case of false starts etc. Only the last token, i. e. the one supposed to be error-free — or at least as good as possible, was taken for further processing. 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.

Three experienced labellers (two of them native speakers of AE, the third one a non-native phonetician who has been living in the US for more than ten years) annotated suprasegmental phenomena such as position of phrase boundaries and accents, WAP if deviant from the correct lexical representation, and non-nativeness of several aspects such as intelligibility on a three-stage rating scale. In the present paper, we want to concentrate solely on the automatic recognition of erroneous placement of WAP. Table 1 displays the frequencies of syllables for types and tokens in the whole database recorded, broken down into numbers of syllables. For obvious reasons, one-syllable words will not be dealt with in this paper.

So far, 24 speakers have been annotated by at least one of the labellers; we want to use solely those 14 speakers which have been annotated by all 3 labellers. These are the 'top ten' words with the highest percentage of erroneous WAP: *'Arabic* (81%), *peni'cillin* (63%), *'Arkansas* (63%), *'lunatic*

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the project *C-AuDiT - Computerunterstütztes Aussprache- und Dialog-Training* under Grant 01IS07014B . The responsibility lies with the authors.

[1] In the following, we simply denote primary word accent position with an apostrophe before the accentuated syllable.

Table 1: *word forms: no. of syllables for types (724) vs. tokens (1609).*

no. syll.	1	2	3	4	5	6	7
# types	318	268	92	29	13	3	1
# tokens	1047	399	108	38	13	3	1
% wrong WAP	0.0	1.9	4.9	5.4	3.2	2.6	0.0

(58%), *sub'ject* (58%), *'discount* (54%), *com'ponents* (50%), *pho'tographer* (46%), *inhu'mane* (46%), and *su'perb* (42%). There seem to be interesting differences between speakers of different L1 but reliable interpretations can only be given when all speakers have been fully annotated.

3. Prosodic Features

The most plausible domain for WA is the syllable level. Later we will see, however, that it can be beneficial to use also features from the word level. We therefore use a feature extraction module that can be applied to arbitrary units of speech. An additional advantage is that the description can be kept generic and short. These units can be as large as phrases or as small as voiced/unvoiced segments; in our case, the units are syllables (across words) and whole words (across phrases). The required segmentation is generated by a forced alignment using a speech recognizer; simpler methods such as a voiced/unvoiced-detector can be used as well [11].

Throughout this paper, it is assumed that the spoken word sequence is identical with the utterance the speaker had to read; utterances where one of the labellers indicated a reading error were disregarded. Using a cross-word tri-phone speech recognizer, a segmentation of the utterance into phonemes and pauses is obtained by a forced alignment. For each utterance, the DC is removed and the maximal amplitude of the signal is normalized. Some of the energy and duration based features described in the following are normalized versions of a quantity, e. g. the duration of a word divided by the average duration of that specific word. The statistics necessary for these normalization measures can be estimated using forced alignment on arbitrary speech material; in our case, the recordings of the four reference speakers are used. Note that this process is text-*independent*: the statistics of unobserved words are estimated from their syllables, and the statistics of unobserved syllables are estimated from their phonemes.

It is still an open question which prosodic features are relevant for different classification problems, and how the different features are interrelated. We try therefore to be as exhaustive as possible and use a highly redundant feature set leaving it to the statistical classifier to find out the relevant features and the optimal weighting of them. However, the procedure is based on knowledge and not on brute force. Many relevant prosodic features are extracted from different context windows with the size of two units before, i. e. contexts -2 and -1, and two units after, i. e. contexts 1 and 2 in Table 2, around the current unit, namely context 0 in Table 2; by that, we use so to speak a 'prosodic 5-gram'. A full account of the strategy for the feature selection is beyond the scope of this paper; details and further references are given in [12].

Table 2 shows the 104 prosodic features and their context. **DurTauLoc** is a local estimate of the speaker-dependent average duration, **EnTauLoc** is a local estimate of the speaker-dependent average energy, and **F0MeanGlob** is the average fundamental frequency. These as well as **RateOfSpeech** are estimated from a window of 15 units (or less, if the utterance is shorter). The other features are abbreviated as follows: **duration features 'Dur':** absolute (Abs) and normalised (Norm); this normalisation is based on duration statistics and on DurTauLoc; **energy features 'En':** regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalised (Norm) values; the normalisation is based on energy statistics and on EnTauLoc; **F0 features 'F0':** regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis;[2] all F0 features are logarithmised and normalised as to the mean value F0MeanGlob; **length of pauses 'Pause':** silent pause before (Pause-before) and after (Pause-after).

Note that these 104 features do not necessarily represent *the* optimal feature set; this could only be obtained by reducing a much larger set to those features which prove to be relevant for the actual task, but in our experience, the effort needed to find the optimal set normally does not pay off in terms of classification performance, cf. [13, 9].

4. Automatic classification

As we confine our analysis to one primary accent and leave aside the three-class problem primary/secondary/no accent, recognition of WAP means just determining the syllable that bears the main, primary accent. For a word with k syllables, we formulate this as a k-class classification problem, with the WAP as the target class. Using linear discriminant analysis (LDA), we build a separate statistical model for each syllable count k, describing the conditional density for the observed word $\boldsymbol{c}$ to have its primary accent at the i-th syllable: $p(\boldsymbol{c}|i,k)$. A priori probabilities are not used at that point as we do not want to favour any (wrong or correct) accent positions. [3]

As acoustic observations $\boldsymbol{c}$, we use the prosodic features set described in Section 3. In the first setup, we compute feature vectors for all syllables. Concatenated, they yield a vector of dimension $k * 104$ for the observed word. Alternatively, we just use the 104-dimensional feature vector computed from the whole word. Eventually we combine both, resulting in $(k+1) * 104$ features being input to the LDA classifier.

Our application context CAPT allows a special optimization. As the contents of a language course are known and prepared in advance, it is feasible to record some few native speakers reading the material, such as our four reference speakers. We utilize this by augmenting the observation $\boldsymbol{c}$ with the reference speakers' *average* observation of the current word. Both observations are concatenated, doubling the number of features used for classification. This process complements the text-independent normalization measures applied during feature extraction (cf. Section 3) and gives the classifier information about the specific current word in the specific current context. Note that the reference speakers need not be annotated for this.

Of particular interest is often not so much the actually realized WAP, but just whether it is the correct one or not. The

[2] These position features are measured in msec.; strictly speaking, they are therefore rather duration features.

[3] In many similar experiments, LDA has proven competitive to more sophisticated classifiers; moreover, it is fast, reliable and robust. Being a statistical classifier, LDA allows for the straightforward derivation of confidences.

Table 2: *104 prosodic features and their context. The features are based on duration (Dur), energy (En), pitch (F0) and pauses. Depending on the mode of analysis, the unit is either word or syllable. Bullets indicate that the features to the left are computed for these context(s) given in columns 2–6. The curly brackets indicate that all the features displayed in these three rows are computed for all contexts in the three rows in columns 2–6.*

features for the actual unit '0' computed from a context of up to two units to the left and right	**context size** -2	-1	0	1	2
DurTauLoc; EnTauLoc; F0MeanGlob; RateOfSpeech			•		
Dur: Norm, Abs		•	•	•	
En: RegCoeff, MseReg, Mean, Abs, Norm	•		•		
F0: RegCoeff, MseReg, Mean		•		•	
En: Max, MaxPos		•	•	•	
F0: Max, MaxPos, Min, MinPos		•	•	•	
Pause-before, F0: Off, Offpos		•	•		
Pause-after, F0: On, Onpos			•	•	

statistical classification approach yields a straightforward confidence measure: the conditional probability of the canonic (i. e. correct) accent position $i^{\text{can.}}$:

$$P(\boldsymbol{c}|i^{\text{can.}},k) = \frac{p(\boldsymbol{c}|i^{\text{can.}},k)}{\sum_{j=1}^{k} p(\boldsymbol{c}|j,k)}. \quad (1)$$

5. Experiments and Results

As mentioned above, we use the subset of 14 speakers that are finished by all three labellers (set "NonN"). For acquiring as good a ground truth as possible for training and evaluation of our system, we merge all available annotations for a speaker; we then simply perform a majority voting for (primary) WAP. We train and evaluate our system using the NonN set in a leave-one-speaker-out (LOSO) cross-validation. As an alternative, we use the four reference speakers (set "Native") for training and the whole NonN set for evaluation. Assuming error-free, canonic realizations, this is an attractive alternative because these speakers do not have to be annotated. This is a big advantage, e. g. when transferring the system to other languages. For comparison, we also evaluate the Native set as an "upper baseline" in a LOSO evaluation.

For comparing the performance of our automatic system with human performance, we estimate the inter-rater agreement of the labellers: we assume one labeller as the ground truth and another as the 'recognition hypothesis' in turn and average the resulting evaluation figures. We also compute the average agreement of a hypothetic 'coward' labeller — one who never deviates from the canonic word accent — with each labeller.

In order to test the different configurations of acoustic features, we first apply our approach to predict WAP. Table 3 lists the percentage of words with correctly predicted WAP, for different train/test setups. Note that the target WAP is the canonic WAP in the case of the LOSO evaluation on the Native set (column 4 of Table 3); otherwise (last two columns), it is the WAP given in the merged annotation — indicating that the speaker has made a mistake by deviating from the canonic position. As can be seen from the figures in the Table, the feature vectors computed from all syllables (bullets in column 1) perform clearly better than the word-based feature vector (bullets in column 2), e. g. 85.6% vs. 69.5 % in rows 1 and 2, last column. Combining both further improves results slightly (e. g. to 86.7 % in row 3, last column). Augmenting the features with the average reference speakers' data as described in Section 4 is also effective, e. g. 92.3 % vs. 86.7 % in rows 3 and 6, last column. The native speakers seem to be a good training set for a mismatched test condition: when training with Native and testing on NonN (last column), performance is similar to the LOSO evaluation on NonN (last three rows, column 5).

The labellers' score for this task — identifying WAP — is 94.6 %, while the coward gets 95.7 % (not contained in Table 3). Apparently, the annotation task is so hard even for the labellers that "daring" to deviate from the canonic WAP costs precision on average. Of course, speakers deviating from the canonic WAP from time to time are vital for having enough training data for recognizing erroneous WAP, which is the task we will look at next. For this 2-class problem *accent position is wrong or not*, the labellers have the following performance: 34.9 % true positive rate (TPR) at 2.7 % false positive rate (FPR). Obviously, the coward strategy here yields 0 % TPR at 0 % FPR.

For automatically recognizing whether WAP is erroneous, we could compute the most likely WAP as above and decide for a mistake if it deviates from the canonic position. A more direct way of looking at the problem is using the confidence for the *canonic* WAP according to (1). We decide for mistake if $P(\boldsymbol{c}|i^{\text{can.}},k) < \theta$. Varying the threshold $\theta \in [0;1]$, we can adjust the system to a desired hit rate (TPR) or false alarm rate (FPR). To improve the system's performance, we optionally use a list of likely error candidates, and only consider those words for mistakes that are included in that selection. Such a list could be compiled manually when designing a language course, based on the experience of the tutors; in our case, we use 12 annotations of 10 speakers that are not yet labelled by all three labellers (and thus not contained in the NonN set). All words that have been annotated by at least one labeller in at least one speaker's recordings have been included in the selection.

Table 3 suggest that employing the reference speakers is beneficial for WAP *recognition*; evaluations which are not reported here showed, however, that this is not the case for the two-class problem *WAP wrong or not*. (Due to the low frequencies of erroneous WAP in the training material, the data of the reference speakers might give the classifier the opportunity to mimick the coward's strategy of predicting just the canonic WAP.) Therefore, we use the feature configuration in row 3 of Table 3 (word and syllable level, but no reference features) for classifying erroneous WAP.

The results for different setups are given in Figure 1 as Receiver-Operator-Curves (ROC). It can be seen that training with the annotated, non-native data yields slightly better

Table 3: *Rate of correctly predicted word accent positions (WAP) for different setups. The bullets in the first three columns denote which of the three feature configurations described in Section 4 are used. Last three columns: % correct WAP.*

using level		using	Native/	NonN/	Native/
syllable	word	ref. spks.	Native	NonN	NonN
•			89.8	85.6	85.6
	•		81.3	72.9	69.5
•	•		91.0	87.6	86.7
•		•	92.7	90.8	90.6
	•	•	90.8	88.2	88.6
•	•	•	94.8	92.0	92.3

performance (at least for a small FPR, the part of the ROC curve that is relevant for applications) than training with the four native reference speakers only: "NonN/NonN" is better than "Native/NonN", as is "NonN/NonN + selection" compared to "Native/NonN + selection". The error candidate selection dramatically improves results, cf. "NonN/NonN + selection" vs. "NonN/NonN" and "Native/NonN + selection" vs. 'Native/NonN". The best system, "NonN/NonN + selection" has a TPR of 34.1 % at the FPR of the labellers (2.7 %) which is nearly as good as the TPR of the labellers (34.9 %). When using no annotated speech data from NonN but only Native for training of the system ("Native/NonN + selection"), it reaches 27.1 % TPR at the labellers' FPR (2.7 %); at the labellers' TPR (34.9 %), a FPR of 3.4 % is achieved, still in the same leage as the labellers in terms of performance.

6. Discussion and Concluding Remarks

Although the performance for recognizing WAP reported in Table 3 meets state-of-the-art requirements, recognizing *erroneous* WAP is obviously a rather difficult problem because the phenomenon is sparse — a real *island* of failure. Additionally, the specific prosody of reading might smear the differences between accentuated and not accentuated (schwa) syllables up to a certain extent, especially in the case of non-native L2 speakers. On the other hand, our system's performance is comparable to the labellers' performance. The fact that training with native data leads only to a moderate loss is important for the application of such a system: these can be obtained at very low costs and do not have to be annotated.

For the eventual application, it seems to be more promising not to decide in favour of any hard decision which then is communicated to the learner, but to give implicit corrective feedback by subsequent pronunciation exercises focusing on the difficult constellations.

7. References

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

[2] C. Cucchiarini, A. Neri, F. de Wet, and H. Strik, "ASR-based pronunciation training: Scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners," in *Proc. Interspeech*, Antwerp, 2007, pp. 2181–2184.

[3] 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: Mouton de Gruyter, 2002, pp. 515–546.

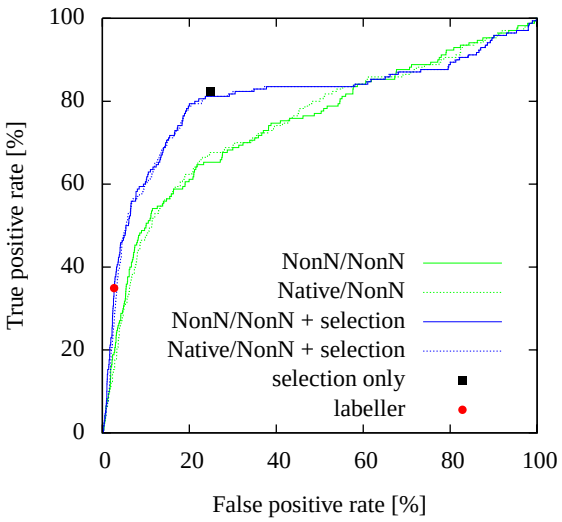

Figure 1: *ROC curves for detecting erroneous WAP, for different setups. "NonN/NonN" refers to a LOSO evaluation on the NonN set; "Native/NonN" to training on Native and testing on NonN. "+ selection" indicates that only words in the list of likely error candidates are considered for errors; "selection only" refers to the strategy of classifying all words as wrong that are in that list. "labeller" designates the average performance of one labeller.*

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

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

[6] J. Tepperman and S. Narayanan, "Better Nonnative Intonation Scores through Prosodic Theory," in *Proc. Interspeech*, Brisbane, 2008, pp. 1813–1816.

[7] P. Lieberman, "Some Acoustic Correlates of Word Stress in American English," *JASA*, vol. 32, pp. 451–454, 1960.

[8] M. Beckman, *Stress and Non–stress Accent*. Dordrecht: Foris Publications, 1986.

[9] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Boiling down Prosody for the Classification of Boundaries and Accents in German and English," in *In Proc. Eurospeech*, Aalborg, 2001, pp. 2781–2784.

[10] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts Prominence; Fundamental Frequency lends little," *Journal of The Acoustical Society of America*, vol. 11, pp. 1038–1054, 2005.

[11] A. Maier, F. Hönig, V. Zeissler, A. Batliner, E. Körner, N. Yamanaka, P. Ackermann, and E. Nöth, "A Language-Independent Feature Set for the Automatic Evaluation of Prosody," in *Proc. Interspeech*, Brighton, 2009, to appear.

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

[13] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Prosodic Feature Evaluation: Brute Force or Well Designed?" in *Proc. ICPHS*, San Francisco, 1999, pp. 2315–2318.