A language-independent feature set for the automatic evaluation of prosody

Andreas Maier, Florian Hönig, Viktor Zeißler, Anton Batliner, E. Körner, N. Yamanaka, P. Ackermann, Elmar Nöth

Angaben zur Veröffentlichung / Publication details:

Maier, Andreas, Florian Hönig, Viktor Zeißler, Anton Batliner, E. Körner, N. Yamanaka, P. Ackermann, and Elmar Nöth. 2009. "A language-independent feature set for the automatic evaluation of prosody." In *INTERSPEECH 2009 - 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, edited by Maria Uther, 600–603. Red Hook, NY: Curran Associates, Inc. https://www.isca-archive.org/interspeech_2009/maier09_interspeech.html.

A Language-Independent Feature Set for the Automatic Evaluation of Prosody

A. Maier[1], *F. Hönig*[1], *V. Zeissler*[1], *A. Batliner*[1], *E. Körner*[2], *N. Yamanaka*[2], *P. Ackermann*[2], *E. Nöth*[1]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2] Lehrstuhl für Japanologie, Universität Erlangen-Nürnberg, Germany

Andreas.Maier@cs.fau.de

Abstract

In second language learning, the correct use of prosody plays a vital role. Therefore, an automatic method to evaluate the *naturalness* of the prosody of a speaker is desirable. We present a novel method to model prosody independently of the text and thus independently of the language as well. For this purpose, the voiced and unvoiced speech segments are extracted and a 187-dimensional feature vector is computed for each voiced segment. This approach is compared to word based prosodic features on a German text passage. Both are confronted with the perceptive evaluation of two native speakers of German. The word-based feature set yielded correlations of up to 0.92 while the text-independent feature set yielded 0.88. This is in the same range as the inter-rater correlation with 0.88. Furthermore, the text-independent features were computed for a Japanese translation of the passage which was also rated by two native speakers of Japanese. Again, the correlation between the automatic system and the human perception of the naturalness was high with 0.83 and not significantly lower than the inter-rater correlation of 0.92.

Index Terms: learning systems, speech analysis, feature extraction.

1. Introduction

Prosody, i. e., the information contained in speech that goes beyond the spoken words, is very important for communication between human beings. It is for example used to disambiguate the meaning of an utterance that is ambiguous given the literal content only. Prosody can also indicate irony and convey mood, affect or emotion. Last but not least, the skillfulness of a speaker is reflected in his or her prosody — mostly involuntarily so. Often non-native speakers exhibit unnatural prosody that makes the speech difficult to understand and listen to. Therefore, prosody plays an important role for acquiring a foreign language.

When learning a foreign language in class, time per individual learner is too short for training the pronunciation of single words efficiently. Training the prosody of the foreign language, however, is often neglected fully. This is also the case for today's computer-assisted language learning systems: although some systems provide automatic pronunciation scoring for single words, prosodic phenomena are mostly ignored [1]. It is therefore desirable to have an automatic method for evaluating the quality of the speaker's prosody. Apart from allowing the learner to train prosody at home, it could also be applied in entry-level or assessment tests.

In this paper, we propose a system to automatically evaluate a speaker's prosody. For evaluation, we concentrate on one specific aspect of speech: the *naturalness* of a speaker's prosody. For the applicability of automatic speech assessment methods, it is very convenient if the system is independent of the learner's specific pair of native (L1) and foreign language (L2), because then no model has to be built for every L1/L2-combination (possibly quadratic in the number of languages) but only for every L2-language in question. Our approach meets this constraint by utilizing only an automatic speech recognition system trained on (native) L2 speech. A text-*independent* approach using features based on voiced segments works even completely without speech recognition.

The paper is organized as follows. Section 2 presents the collection and annotation of the speech data used for training and evaluation. Section 3.1 introduces the features based on a speech recognizer, Section 3.2 the text-independent features. The rest of that Section covers the tools used for estimating the naturalness of the speaker's prosody from the computed features. Section 4 presents results, which are discussed in Section 5. Section 6 closes with a summary.

2. Data

The recorded data consists of both the German and the Japanese version of the same text passage from the book “The Little Prince” by Antoine de Saint-Exupéry. It was considered appropriate for the aim of targeting prosodic features, due to its balanced difficulty and literacy requirements.

26 people were recorded altogether, aged between 19 and 53 years: 8 German female, 5 German males, 9 Japanese females, and 4 Japanese males. All subjects were able to speak German and Japanese. The task was to read the German and the Japanese version of the text. All subjects were provided an opportunity to read and practice the texts beforehand in order to be able to concentrate on reading as fluently and naturally as possible during recording.

The German text covers 183 words and took German subjects about 1 min. 15 sec. to read in average, whereas Japanese native speakers needed about 30 seconds more to complete the text. The Japanese version of the text took Japanese subjects about 1 min. 45 sec. to read, German native speakers needed about 1 minute more to complete.

The microphone was carefully positioned near the reader's mouth, assuring avoidance of exhaled air noise as well as enabling minimization of possible background noise. We recorded at a sample rate of 16 kHz with 16 bit.

The collected data was independently evaluated by two native speakers of Japanese and two native speakers of German, who separately listened to each record of each subject and evaluated the naturalness of the speaker's prosody on a scale from 1 to 5. All raters had experience in teaching their own native language. In order to form a reference for the automatic evaluation system the mean of both raters was computed.

3. Methods

3.1. Word-based Prosodic Features

In order to compute prosodic features, the output of a word recognition system in addition to the speech signal is required. In this case, the time-alignment with the Viterbi algorithm of our recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used to calculate our prosodic word features [2].

First so called base-features are extracted from the speech signal. These are the energy, the fundamental frequency (F_0) after [3], and the voiced and unvoiced segments of the signal. In a second step, the actual structured prosodic features are computed to model the prosodic properties of the speech signal. For each word we extract 21 prosodic features. These features model F_0, energy and duration, e.g. maximum of the F_0. Fig. 1 shows examples of the F_0 features. In addition, 16 global prosodic features for the whole utterance are calculated. They cover mean and standard deviation for jitter and for shimmer, the number, length and maximum length both for voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of length of voiced sections to the length of the signal and the same for unvoiced sections. The last global feature is the variance of the fundamental frequency F_0. In order to evaluate the speech, we calculate the average, the maximum, the minimum, and the variance of the 37 turn- and word-based features for the whole text to be read. Thus we get 148 features for the whole text.

Fig. 1 shows examples of the F_0 features. The mean values F0MeanGlobalWord are computed for a window of 15 words (or less if the utterance is shorter) [4, 5] so they are regarded as turn-level features here.

3.2. Text-Independent Prosodic Features

The text-independent prosodic features are computed without using a speech recognizer or some kind of forced time-alignment algorithm. As the name indicates, they are completely independent from the textual content of the spoken sentence. In fact, these features can be used for prosodic analysis of any language without any algorithmic changes.

The computational procedure is very similar to the computation of the word-based features. The prosodic base-features including energy, voiced and unvoiced segments, F_0 and pitch periods are extracted from the speech signal and then normalized using log-scaling and mean subtraction techniques. The computation of the structured prosodic features at the next step requires a segmentation: here we take the voiced segments instead of words. We merge the adjacent segments when they are separated by less than 50 msec and interpolate the corresponding F_0 contour to make the segmentation more robust. In addition to the single segments we also use the so-called *context segments* consisting of two adjacent segments merged together.

The prosodic features we compute on the voiced segments differ slightly from the word-based feature set. There are no on- and offsets (and their positions) for F_0 features because they are identical to the segment boundaries. Instead, we additionally compute a spectral description of the F_0 and energy contours using the absolute values of the first 10 FFT-coefficients of the 128-point FFT-window centered over the current segment which corresponds to a 1.28 second window at our frame rate of 10 msec. These FFT-features are computed only for the single segments. For every segment position, including the current single segment and its context segments, we extract a total of 187 features.

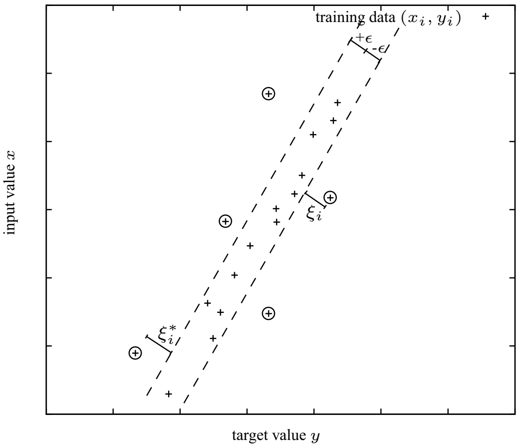

Figure 2: *Support Vector regression finds a function that has at most ϵ deviation from the targets y_i. In order to allow deviations larger than ϵ a slack variable ξ_i is introduced again. Note that the support vectors are outside the ϵ-tube.*

The further processing runs similar as for the case of word-based features. The 187 segment features are combined to 187 text-features by computation of the mean value for each feature. These are used for the regression later on.

3.3. Support Vector Regression

In the next step a model to predict the actual target value y_i has to be created. In order to approximate an arbitrary function, *Support Vector Regression* (SVR) [7] can be applied. It's goal is to compute an estimate value $\hat{y_i}$ for each of the N feature vectors $\vec{x_i}$ which deviate at most ϵ from the original target value y_i. This leads to the following equation:

$$\hat{y_i} = \vec{w}^\top \vec{x_i} + b. \tag{1}$$

The variables $\vec{w}$ and b are found by solving the problems

$$y_i - (\vec{w} \cdot \vec{x_i} + b) \leq \epsilon \quad \text{and} \quad (\vec{w} \cdot \vec{x_i} + b) - y_i \leq \epsilon. \tag{2}$$

To allow deviations greater than ϵ, the slack variables ξ_i and ξ_i^* are introduced. Equation 2 changes then to

$$y_i - (\vec{w} \cdot \vec{x_i} + b) \leq \epsilon + \xi_i \quad \text{and} \quad (\vec{w} \cdot \vec{x_i} + b) - y_i \leq \epsilon + \xi_i^*. \tag{3}$$

In order to constrain the type of the vector $\vec{w}$ we postulate *flatness*. One way to achieve this is to minimize it's norm $||\vec{w}||$. We end up with the following minimization problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}||w||^2 + C \sum_i (\xi_i + \xi_i^*) \\ &\text{subject to} && \begin{cases} y_i - (\vec{w} \cdot \vec{x_i} + b) \leq \epsilon + \xi_i \\ (\vec{w} \cdot \vec{x_i} + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{4}$$

As is the case for Support Vector Machines [8], a primal Lagrangian can be formulated introducing Lagrange multipliers α_i, α_i^*, η_i, and η_i^* in order to solve this problem. The Lagrange multipliers η_i and η_i^* are eliminated in the derivation of the primal Lagrangian. According to [7] the constraint $\alpha_i \alpha_i^* = 0$ has to be met. Thus, there can never be a set of variables α_i and

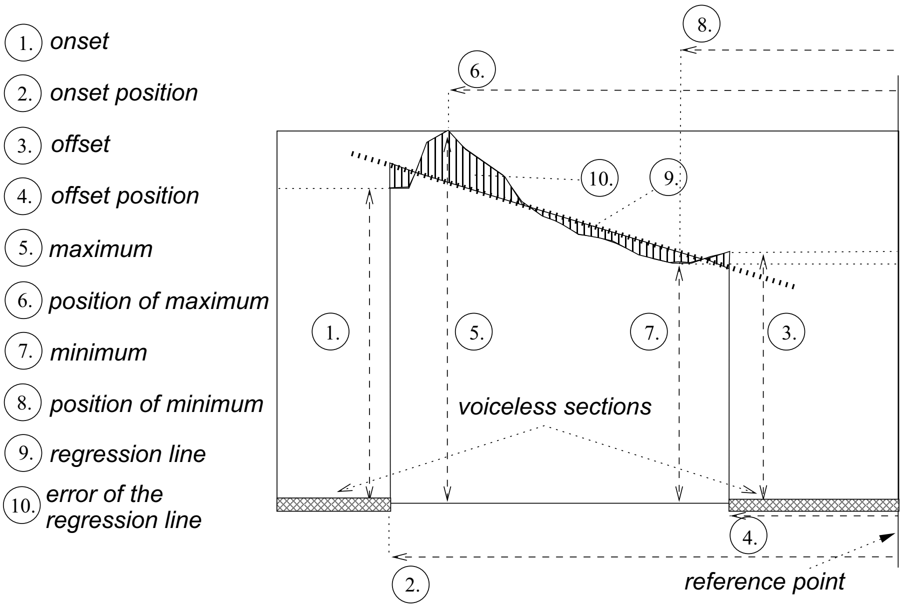

Figure 1: *Computation of prosodic features within one word (after [6])*

α_i^* where both are nonzero at the same time. Furthermore α_i and α_i^* are zero if $|\hat{y_i} - y_i| \leq \epsilon$. Therefore, *support vectors* can only be found outside the ϵ-tube (cf. Figure 2). With the Support Vector Expansion, the prediction of $\hat{y_i}$ from Eq. 1 can be written without the actual weight vector $\vec{w}$:

$$\hat{y_i} = \left[\sum_j (\alpha_j - \alpha_j^*) \vec{x_j} \right]^\top \vec{x_i} + b \qquad (5)$$

Hence, only the Support Vectors have to be stored in order to compute the regression.

3.4. Feature Selection

In this work *Correlation-based Feature Subset* (CFS) selection combined with a best-first search as provided by [9] is applied to select an optimal subset of the full feature set. The idea behind the CFS selection algorithm is to compute the correlation of a composite variable $\mathcal{X}^{\mathcal{S}}$ to an outside variable $\mathcal{Y}$ as the criterion for the quality. In [10, p.182] a formulation of this correlation as a composition of the inter-correlations $r_{\mathcal{Y}x_i^{\mathcal{S}}}$ between the target variable $\mathcal{Y}$ and the $N_{\mathcal{S}}$ individual features $x_i^{\mathcal{S}}$ and the intra-correlations $r_{x_i^{\mathcal{S}} x_j^{\mathcal{S}}}$ is found:

$$r_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}} = \frac{N_{\mathcal{S}} \overline{r_{\mathcal{Y}x_i^{\mathcal{S}}}}}{\sqrt{N_{\mathcal{S}} + N_{\mathcal{S}}(N_{\mathcal{S}} - 1)\overline{r_{x_i^{\mathcal{S}} x_j^{\mathcal{S}}}}}} = G_{\text{CFS}}^{\mathcal{S}} \qquad (6)$$

$\overline{r}$ denotes the mean of the respective correlations. In [11] Eq. 6 is used to create a fast and efficient algorithm to select features which have a good correlation with the target variable. The computation is very efficient, since the correlations between all variables just have to be computed once. After their computation the single correlations are stored in a lookup-table which allows fast and easy access to the values.

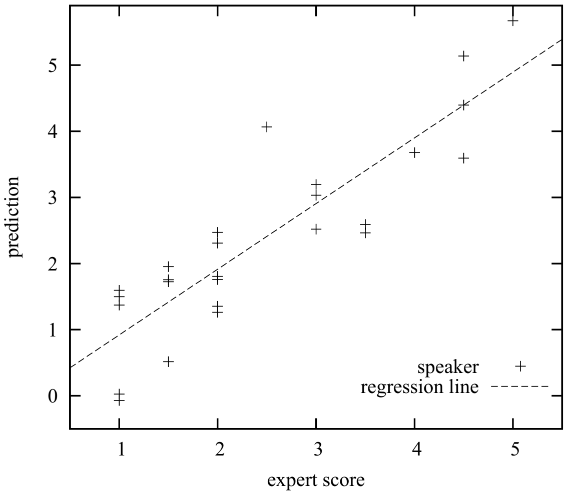

Figure 3: *Correlation between the experts' opinion and the prediction of the LOO system for the German language and the text-independent features (r* $= 0.88$*).*

4. Experiments and Results

All experiments were conducted in a leave-one-speaker-out (LOO) manner. Table 1 reports Pearson's correlation coefficients between the raters and the automatic system. All reported correlations are significant at $p < 0.01$. Spearman's correlation coefficient was also investigated. Since the data was normally distributed both coefficients were in the same range. Hence, only Pearson's correlation is reported. The reported results were obtained with a linear SVM kernel. The use of higher polynomials did not yield any improvements.

The inter-rater correlation was very high with 0.88 for the German version of the text and 0.92 for the Japanese version. Investigation of the agreement between the two raters of each

Table 1: *Correlations between the automatic evaluation system and the human raters in comparison to the inter-rater correlation*

language	inter-	word-based		text-independent	
	rater	SVR	SVR (CFS)	SVR	SVR (CFS)
German	**0.88**	0.89	**0.92**	**0.88**	0.75
Japanese	**0.92**	-	-	0.76	**0.83**

language with the weighted Kappa [12] yielded coefficients of $\kappa = 0.72$ for German and $\kappa = 0.82$ for Japanese which corresponds to a high agreement in both cases.

With the word-based prosodic features, a correlation of 0.89 could be achieved. Feature selection in each leave-one-out iteration could improve this even further to 0.92. A word-based prosodic evaluation was not performed on the Japanese version of the text because segmentation into words as in western languages is not straightforward in Japanese.

The text-independent prosodic features also yielded a high performance on the German speech data with a correlation of 0.88 (cf. Figure 3) without feature selection and 0.75 with CFS selection. The correlations between the perceptive evaluation and the automatic system on the Japanese data was also comparable: A correlation of 0.76 was achieved without CFS selection and 0.83 with feature selection. A reason for the differences between German and Japanese could be that the feature set was originally designed and evaluated with German speech data only. Hence, some of the features might not be meaningful for the Japanese data and should therefore be excluded.

5. Discussion

The perceptive evaluation of the native speakers was very consistent in German and in Japanese. Hence, the raters could determine speakers with natural prosody easily and their agreement on this feature was high. Therefore, the mean of the raters' opinion is suitable to train an automatic evaluation system.

The results of the word-based evaluation system were in the same range as the human evaluation. In fact the correlation was even slightly higher than the inter-rater correlation in two of four cases. No significant difference was found ($p > 0.05$), i.e., the automatic evaluation is as reliable as the ones of the experts. Significance testing with the u-test was performed after [13].

On the German data the results of the text-independent prosodic features were slightly worse than the word-based prosodic features. However, no significant difference between the inter-rater correlation, the word-based evaluation system, and the text-independent evaluation system was found ($p > 0.05$). Hence, the performance of both evaluation systems and the perceptive evaluation can be regarded as comparable.

With the text-independent system, an automatic evaluation of the prosody of the Japanese data could also be performed. It's performance was worse than the experiments with the German data. A significance test between the inter-rater correlation and the SVR system showed that there was no significant difference between both ($p > 0.05$). Thus, also the automatic evaluation of the Japanese automatic system can be regarded as comparable to the perceptive evaluation.

6. Summary

Our novel approach is able to evaluate the *naturalness* of the prosody of a speaker. One variant of the system is independent of the language, because it obtains the time alignment information automatically from the structure of the speech data using voiced and unvoiced segments. Hence, there is no speech recognition required and the system can be applied as is to any other language.

On the German data it could be shown that these features allow a comparable performance to word-based prosodic features. The accuracy of the system was in the same range as the perceptive evaluation of native speakers of the respective language. For German the inter-rater correlation was 0.88 while the system's performance also was 0.88 and for Japanese the inter-rater correlation was 0.92 while the system had a correlation of 0.83 to the mean of the human raters.

7. References

[1] C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth, "Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. Hawaii, USA: IEEE Computer Society Press, 2007, pp. 197–200.

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

[3] P. Bagshaw, S. Hiller, and M. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*. Berlin, Germany: ISCA, 1993, pp. 1003–1006. [Online]. Available: citeseer.ist.psu.edu/169670.html

[4] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Prosodic Feature Evaluation: Brute Force or Well Designed?" in *Proc. of the 14th Intl. Congress of Phonetic Sciences (ICPhS)*, vol. 3, San Francisco, USA, 1999, pp. 2315–2318.

[5] 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 *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, vol. 4. Aalborg, Denmark: ISCA, 2001, pp. 2781–2784.

[6] A. Kießling, *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*, ser. Berichte aus der Informatik. Aachen, Germany: Shaker, 1997.

[7] A. Smola and B. Schölkopf, "A tutorial on support vector regression," Royal Holloway University of London, Tech. Rep., 1998, nC2-TR-1998-030.

[8] B. Schölkopf, "Support vector learning," Ph.D. dissertation, Technische Universität Berlin, Germany, 1997.

[9] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Fransisco, CA, USA: Morgan Kaufmann, 2005.

[10] E. Ghiselli, *Theory of Psychological Measurement*. New York, USA: McGraw-Hill Book Company, 1964.

[11] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[12] M. Davies and J. Fleiss, "Measuring agreement for multinomial data," *Biometrics*, vol. 38, no. 4, pp. 1047–1051, 1982.

[13] K. Stange, *Angewandte Statistik II*. Berlin, Heidelberg, Germany: Springer Verlag, 1971.