ISCA Archive
http://www.isca-speech.org/archive

Second Language Studies – Acquisition, Learning, Education and Technology
Tokyo, Japan
September 22-24, 2010

How Many Labellers? Modelling Inter-Labeller Agreement and System Performance for the Automatic Assessment of Non-Native Prosody

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

[1]Pattern Recognition Lab, Universität Erlangen-Nürnberg, Germany
[2]digital publishing, München, Germany
{hoenig,batliner}@informatik.uni-erlangen.de

Abstract

On a database of non-native English productions annotated by 60 native English speakers as for their quality w. r. t. intelligibility, non-native accent, melody and rhythm, we study how inter-labeller correlation and performance of a regression system change when varying the number of labellers used for training. This depends highly on the difficulty of the labelling task, the features used by the regression system and the type of regression used. We propose a model that parametrises these dependencies and is able to predict the system's performance when increasing the number of labellers. This can provide a valuable basis for decision-making when trying to improve an existing regression system as efficiently as possible. We show the plausibility of our approach by experimental evaluation.
Index Terms: non-native prosody, speech melody, rhythm, crowdsourcing, inter-labeller agreement, regression system, performance model

1. Introduction

Non-native prosodic, especially rhythmic traits are a main source for low intelligibility of the speech of non-native L2 speakers of English – and any other language. To assess such traits automatically, we normally need data that are annotated as for the degree of deviation from native prosody, serving as 'reference' or 'ground truth' for training automatic procedures such as classifiers or regression. Note that the following statements can be conceived as generic, valid for any annotation task, not only for prosodic assessment which is the topic of this article: apart from the speech data that should be annotated – type, size, sub-samples such as male/female, degree of proficiency, etc. – the main alternatives to be chosen from is a choice between experts and 'naive' subjects for annotation and/or perceptive evaluation, and the decision on how many people to employ. 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 [1]. 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 [1]. Experts are rare and more expensive than 'naive' subjects; moreover, they may be biased in some way towards their own theoretical preferences. Naive subjects are less expensive, thus more of them can be employed, they are less biased, but care has to be taken that the task is well-defined. Thus normally, less experts are employed than naive subjects. How many to 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 5 or more labellers, we are more safe when establishing ordinal judgements, based on the average score of all annotators. Intuitively, around 10 is a good figure; more than 20 are employed rather rarely. However, to our knowledge, it has not been investigated systematically yet how the number of labellers influences the performance of automatic procedures.

The variability between labellers can be traced back to at least two main factors: first, speaker-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.

This paper is a continuation of [2] where we assessed the same task – however, always based on the full set of human labellers. Thus, in [2], the question was how good we are with different input features when we use all information we do have, in the present paper, the question is how does performance change when systematically varying the number of labellers.

2. Material and human assessment

We recorded 55 English L2 speakers: 25 German, 10 French, 10 Spanish, and 10 Italian speakers.They had to read aloud 329 utterances shown on the screen display of an automated recording software.The data to be recorded are described fully in [2]. Based on annotations of three experienced labellers [3], we defined a subset of the five sentences that were judged as 'prosodically most error-prone for L2 speakers of English', cf. [2].

For annotation, a perception experiment was conducted for scoring intelligibility, non-native accent, perceived L1, melody and rhythm, using the tool PEAKS [4]. 20 native American English, 19 native British English, and 21 native Scottish English speakers with normal hearing abilities judged each sentence in random order. As shown in [2], there are no real differences between judgements from these three varieties of English. Thus, all 60 labellers are lumped together. We only deal with the answers to the melody question in this paper (THIS SENTENCE'S MELODY SOUNDS: (1) *normal* (2) *acceptable, but not perfectly normal* (3) *slightly unusual* (4) *unusual* (5) *very unusual*). The labels on the Likert scales were averaged over all sentences of a speaker to get a single score for each criterion.

3. Features

After segmenting the recordings with forced alignment of the target utterance using a cross-word triphone HMM speech recognition system, we automatically compute a large number of features measuring different prosodic traits on speaker level (a more detailed description is given in [2]):

Speech Rate Measures: 6 features *SR* describing the rate of syllables, stressed syllables and vocalic segments.

Isochrony Features: 12 features *Iso* capturing distances between stressed and between unstressed syllables and the standard deviations of those distances, in order to capture possible isochrony properties [5].

Variability Indices: Following [6], we identify vocalic and consonantal segments and calculate the raw Pairwise Variability Index (rPVI) which is defined as the absolute difference in duration of consecutive segments and its normalized version nPVI for vocalic and consonantal segments. We compute 8 speaker-level Pairwise Variability Index features *PVI*.

Global Proportions of Intervals: Following [7], we compute the percentage of vocalic intervals (of the total duration of vocalic and consonantal segments), the standard deviation of the duration of vocalic and consonantal segments, and derive 6 features *GPI* measuring Global Proportions of Intervals.

General-Purpose Prosodic Features: In addition to the specialized features, we apply our comprehensive general-purpose prosody module which has already been successfully applied to diverse problems such as phrase accent and phrase boundary recognition, word accent position classification, and emotion recognition [2]. The features are based on duration, energy, pitch, and pauses, and describe arbitrary units of speech (in our case words, syllables, and nuclei) by 35 features (or 104, if context is included). A more detailed overview of the prosodic features is given in [8]. We use these prosodic features computed over different units and contexts to construct extensions of the *Iso*, *PVI* and *GP* features to form a total of 523 general-purpose prosodic features *Pros*.

Speech Recognition Features Additionally, we use 6 features *WR* describing the accuracy of a free unigram speech recognizer with respect to the target utterances.

4. Modelling Labeller and System Performance

In order to predict the speaker's melody score from the features, we apply multiple linear regression in two setups, which differ in the way dimensionality is reduced before applying regression. In the *PCA regression system*, we apply PCA using the Kaiser-Guttman criterion to select up to a maximum of 40 principal components. In the alternative setup, we apply feature selection (FS) and use the 5 best features resulting from a greedy forward search in a wrapper approach. We refer to this system as the *FS regression system*. We evaluate the performance of the systems in terms of the average Pearson correlation coefficient in a 10-fold, speaker-independent cross-validation.

We denote the *Pearson correlation coefficient* between two random variables A and B by $\rho_{A,B} = \mathrm{Corr}(A,B) = \frac{\mathrm{Cov}(A,B)}{\sigma_A \sigma_B}$. Its estimate computed from samples of A and B is the *sample correlation coefficient* $r(A,B)$. We adopt a very coarse model for the annotations given by the labellers: we do not distinguish between speaker-specific traits and states, and neither account for differently scaled labels nor for the fact that different pairs of labellers have different correlation coefficients. The annotations X_k of the labellers $k = 1, 2, \ldots$ are modelled as jointly normally distributed random variables with $\mathrm{Var}(X_k) = \sigma^2$ and $\mathrm{Cov}(X_i, X_j) = c\sigma^2$ for any pair of labellers $i \neq j$. Consequently, the annotations of two labellers have a Pearson correlation coefficient of c, which reflects how competent the labellers are for the given labelling task.

Combined Annotations formed by linear combination[1] of multiple labellers $X_1, X_2, \ldots, X_N$ are denoted by

$$X^N := \left(\sum_{k=1}^{N} X_k\right) \Bigg/ \sqrt{\mathrm{Var}\Big(\sum_{k=1}^{N} X_k\Big)}\,,$$

which leads to a natural definition of the (imaginary) "ground truth" labels as $L := \lim_{N\to\infty} X^N$.

What is the correlation coefficient between combined annotations? Let Y^M be a combined annotation formed from labellers $X_{N+1}, X_{N+2}, \ldots, X_{N+M}$, i. e. a group of M labellers *disjunct* from the group that forms X_N. Then we get

$$\mathrm{Corr}(X^N, Y^M) = \frac{c}{\sqrt{\frac{1}{N} + \frac{N-1}{N}c}\sqrt{\frac{1}{M} + \frac{M-1}{M}c}}. \qquad (1)$$

c can be estimated by computing $r(X^N, Y^M)$ on samples of X_k and solving (1) for c. Comparing a combined annotation X^N with the ground truth L yields $\mathrm{Corr}(X^N, L) = \sqrt{c/(\frac{1}{N} + \frac{N-1}{N}c)}$.

We model the labels $\hat{X}^N$ produced by the automatic regression system when trained with X^N as the sum of its training labels and two independent error components:

$$\hat{X}^N = X^N + E_i + E_l(N),$$

with expected values $\mathrm{E}(E_i) = \mathrm{E}(E_l(N)) = 0$ and $\mathrm{Var}(E_i) = e_i$, representing an "internal" error of the system (due to suboptimal input features, parameter estimation from finite sample, violation of model assumptions, etc.). $E_l(N)$ increases with the derivation of the labels from the ground truth (bad training labels are normally harder to predict because they are less consistent with the input features). We choose its variance proportional to the fraction of unexplained variance of X^N with respect to L (which equals $1 - \mathrm{Corr}(X^N, L)^2$), i. e. $\mathrm{Var}(E_l(N)) = e_l(1 - c/(\frac{1}{N} + \frac{N-1}{N}c))$. When training the system with X^N, its output and Y^M correlate as follows:

$$\mathrm{Corr}(\hat{X}^N, Y^M)) = c \Bigg/ \Bigg(\sqrt{\frac{1}{N} + \frac{N-1}{N}c}\sqrt{\frac{1}{M} + \frac{M-1}{M}c}\sqrt{1 + e_i + e_l\Big(1 - \frac{c}{\frac{1}{N} + \frac{N-1}{N}c}\Big)}\Bigg). \qquad (2)$$

e_i and e_l can be estimated by computing $r(\hat{X}^N, Y^M))$ for two different values of N and solving the resulting instances of (2).

Up to here, we expressed dependencies between combined annotations formed from *disjunct* groups of labellers. In the remainder of this section, we will give useful relations for *overlapping* groups of labellers. The correlation of a single labeller's annotation X_1 and the combined annotation X^N from N labellers, *including* X_1, is

$$\mathrm{Corr}(X_1, X^N) = \frac{1 + (N-1)c}{\sqrt{N + N(N-1)c}}, \qquad (3)$$

i. e., c can be computed from an estimate $r(X_1, X^N)$ by

$$c = \frac{N \cdot r(X_1, X^N)^2 - 1}{N - 1}. \qquad (4)$$

When trained with a single labeller's annotation X_1, the regression system output $\hat{X}_1$ and the combined annotation X^N from N labellers, *including* X_1, correlate with

$$\mathrm{Corr}(\hat{X}_1, X^N)) = \frac{1 + (N-1)c}{\sqrt{N + N(N-1)c}\sqrt{1 + e_i + e_l(1-c)}} =: \rho_1. \qquad (5)$$

Training the regression system with the combined annotation X^N from N labellers and testing with the *same* combined an-

[1] For ease of notation, X^N is not shifted to a certain mean or scaled to variance σ^2, as the correlation is independent of shifting and scaling.

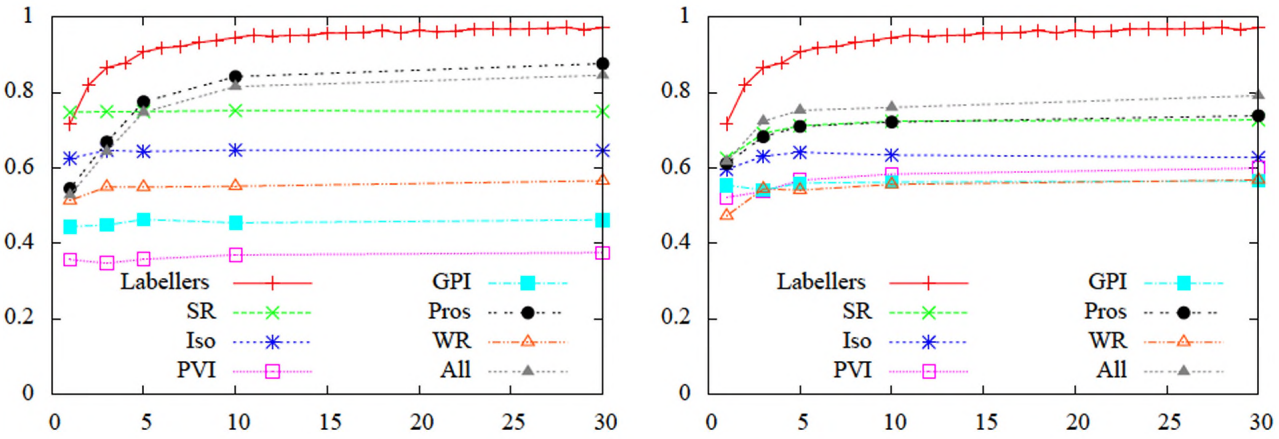

Figure 1: Left: Feature reduction with PCA; Right: Feature Selection. 'Labellers': Estimated labeller correlation $r(X^N, Y^M)$ between average melody annotations of $N = 1 \ldots 30$ (x-axis) and $M = 30$ independent labellers. 'SR', 'Iso', etc.: Estimated correlation $r(\hat{X}^N, Y^M)$ between output of regression system and averaged annotations from $M = 30$ labellers when training with $N = 1, 3, 5, 10, 30$ independent labellers, for the different feature sets 'SR', 'Iso', etc. as input.

notation X^N yields

$$\mathrm{Corr}(\hat{X}^N, X^N) = \frac{1}{\sqrt{1 + e_i + e_l\left(1 - \frac{c}{\frac{1}{N} + \frac{N-1}{N}c}\right)}} =: \rho_N. \quad (6)$$

e_i and e_l can be computed from estimates $r(\hat{X}_1, X^N))$ and $r(\hat{X}^N, X^N)$ of ρ_1 and ρ_N by solving (5) and (6) for

$$e_i = \Big(c(N^2(\rho_1^2 - \rho_1^2\rho_N^2) + N(\rho_1^2\rho_N^2 - \rho_1^2 - \rho_N^2) + \rho_N^2) + N\rho_1^2 - \rho_N^2\Big) / \Big(N(N-1) \cdot c\rho_1^2\rho_N^2\Big), \text{ and} \quad (7)$$

$$e_l = \Big(c(c\rho_N^2(2N - N^2 - 1) + N(N\rho_1^2 - 2\rho_N^2 - \rho_1^2) + 2\rho_N^2 + \rho_1^2) - \rho_N^2\Big) / \Big(N(N-1) \cdot c(c-1) \cdot \rho_1^2\rho_N^2\Big). \quad (8)$$

Note that the performance of the system w. r. t. the ground truth is lower than w. r. t. the annotations, namely

$$\mathrm{Corr}(\hat{X}^N, L) = \sqrt{\frac{c}{(\frac{1}{N} + \frac{N-1}{N}c)}} \cdot \rho_N. \quad (9)$$

Using (9), we can now predict how performance will increase when collecting annotations from more labellers. As N approaches infinity, the performance of the regression system is predicted to approach $\mathrm{Corr}(\hat{L}, L) = \frac{1}{\sqrt{1+e_i}}$.

Summing up, using the parameters c, e_i and e_l, we modelled the correlation between annotations composed from multiple labellers, and the performance of a regression system trained with those composed annotations, depending on the number of labellers involved. We started with annotations formed from *disjunct* groups of labellers, $Corr(X^N, Y^M)$ and $Corr(\hat{X}^N, Y^M)$, and ended up with the more convenient expressions for the case of *overlapping* groups of labellers, $Corr(X_1, X^N)$, $Corr(\hat{X}_1, X^N)$ and $Corr(\hat{X}^N, X^N)$.

5. Experiments and Results

In the following, we experimentally evaluate inter-labeller correlation and performance of the two different regression systems with various input features and annotations when varying the number of labellers. For estimating $\mathrm{Corr}(X^N, Y^M)$, we shuffle and split the 60 labellers into two halves, and compute annotations X^N with $N \leq 30$ from the first half of labellers, and Y^M with $M = 30$ from the second half, and compute $r(X^N, Y^M)$. This process is repeated for 20 random partitions, and the results are averaged.

For estimating $\mathrm{Corr}(\hat{X}^N, Y^M)$, we train the system with X^N computed from the first half of labellers, and compare its outputs[2] with Y^M computed from the second half of labellers, and average $r(\hat{X}^N, Y^M)$ over 20 random partitions.

In Figure 1 the estimated values $r(X^N, Y^M)$ for the correlation between combined annotations from independent groups of $N = 1 \ldots 30$ and $M = 30$ labellers are plotted, and the estimated performance $r(\hat{X}^N, Y^M)$ of the automatic system depending on the number of labellers $N = 1, 3, 5, 10, 30$ used for training, for different input features. Apart from some noise, the inter-labeller correlation ('Labellers' in Figure 1) rises as expected with growing N; the improvement from $r = 0.72$ to $r = 0.97$ as N increases from 1 to 30 is quite notable. The performance of the regression system also rises with growing N, generally speaking (apart from some noise), but obviously the behaviour depends strongly on the used features and the regression system. For example, the PCA system (Figure 1 left) with *SR* features cannot make much use of more labellers: $r = 0.75$ for both $N = 1$ and $N = 30$, while the same system improves dramatically from $r = 0.55$ to 0.88 when using *Pros* features. The behaviour of *SR* and *Pros* is again different for the system using feature selection (Figure 1 right): here, performance rises moderately in both cases, from $r = 0.63$ to $r = 0.73$ (*SR* features) and from $r = 0.61$ to $r = 0.73$ (*Pros* features).

Especially for the generally better performing PCA features (Figure 1, left), the "simple" *SR* features perform best amongst the special-purpose features; *Pros* and *All* are, however, superior, maybe because they can model specifities of the data better – but for doing that, they need more labellers: both sets display a pronounced rising from 1 to approx. 10 labellers, compared to all other feature sets.

Figure 2 shows estimated inter-labeller correlation and estimated system performance for the example of the PCA system

[2] In order not to get optimistic results, we compute all outputs on the unseen test data of each cross-validation fold

using *Pros* features ('Pros' in Figure 1 left; 'System' in Figure 2) along with predictions made by our models using the parameters c, e_i and e_l. For predicting the inter-labeller correlation $\mathrm{Corr}(X^N, Y^M)$ according to (1), we compute $c = 0.52$ from $r(X^N, Y^M) = 0.97$ at $N = 30$. The prediction ('Lab. pred') matches the values estimated from the annotation data ('Labellers') very closely across $N = 1 \ldots 30$. This is remarkable as it is just tuned with one single parameter from the estimate at $N = 30$. This is a strong indication that the coarse model of the labellers adopted is sufficient for our purposes.

In order to predict the system's performance $\mathrm{Corr}(\hat{X}^N, Y^M)$ according to (2) we computed the parameters $e_i = 0.20$ and $e_l = 1.1$ from $r(\hat{X}^N, Y^M)$ at $N = 1$ and $N = 30$. The prediction ('System pred.' in Figure 2) matches the values estimated from experimental evaluation ('System' in Figure 2) relatively closely which makes the model obviously a useful one, e. g. for predicting which performance could maximally be acquired by increasing the number of labellers ('System ∞' in Figure 2).

To give an illustrative example: the model predicts that, given a pair-wise labeller correlation of $c = 0.52$ and that particular PCA regression system using *Pros* features with $e_i = 0.20$ and $e_l = 1.1$, maximal performance $\mathrm{Corr}(\hat{L}, L) = 1/\sqrt{1 + e_i} \approx 0.91$, and using one labeller for training will on average yield 60% (relative) of that upper limit, 5 labellers 85%, 10 labellers 90%, 20 labellers 95%, and 40 labellers 96%. In terms of explained variance, this corresponds to 36% (relative) for one labeller, 72% for 5 labellers, 83% for 10 labellers, 89% for 20 labellers, and 93% for 40 labellers.

In practice, estimating the model parameters a, e_i and e_l by iterating over multiple labeller partitions is cumbersome. We can estimate the parameters more conveniently with the help of (4), (7) and (8). Doing so with $N = 60$, we are still able to predict $\mathrm{Corr}(X^N, Y^M)$ precisely (therefore not shown in Figure 2) and the predictions for $\mathrm{Corr}(\hat{X}^N, Y^M)$ are still reasonably good ('System pred. 2' and 'System ∞ 2' in Figure 2).

6. Discussion and Concluding Remarks

Strictly speaking, it might not be possible to give a general recommendation as for the number of labellers one should hire – it depends on the difficulty of the annotation task, the regression system used, and the accuracy that is needed by the application. But given a working regression system and labels from a non-trivial number of labellers, we can make some educated guesses (see penultimate paragraph of Section 5). As a rule of thumb, the improvement from one to five labellers is marked, and still clearly visible from six to some ten; thus, this might be the region where it definitely pays off to employ more labellers.

As we have shown, the correlation between groups of labellers is very much predictable from the average pairwise correlation, which can conveniently be estimated by comparing each single labeller with all labellers using (3). For predicting the performance of a regression system, however, used input features and used regression system have to be taken into account as well. Our model parametrizes these dependencies and is able to approximately predict performance as a function of the number of labellers. The plausibility of our approach has been demonstrated by experimental evaluation. This model can serve as a valuable basis for decision-making when trying to improve an existing regression system as efficiently as possible (e.g. should one invest money in more labellers or rather try to improve the input features and/or the regression technique). An interesting direction of future research is to incorporate the sample size into our model.

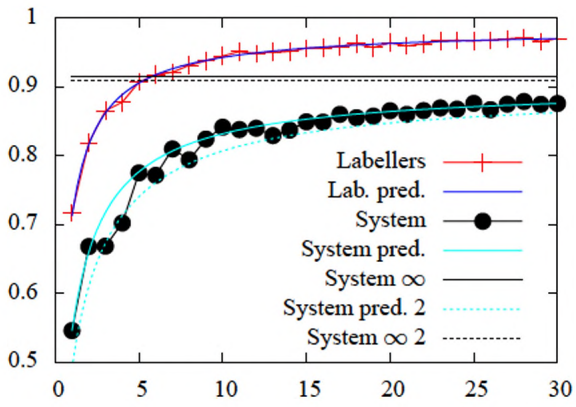

Figure 2: 'Labellers': *Estimated* inter-labeller correlation $r(X^N, X^M)$ for $N = 1 \ldots 30$ (x-axis) and $M = 30$. 'Lab. pred.' (almost coincides with 'Labellers'): *Predicted* inter-labeller correlation $\mathrm{Corr}(X^N, Y^M)$ using Eq. (1); 'System': *Estimated* correlation $r(\hat{X}^N, Y^M)$ between output of regression system (PCA, *Pros* features) and averaged annotations from $M = 30$ labellers when training with $N = 1 \ldots 30$ labellers. 'System pred.': *Predicted* correlation $\mathrm{Corr}(\hat{X}^N, Y^M)$ between regression system and labellers using Eq. (2). 'System ∞': Predicted correlation between regression system and labellers when training with $N \to \infty$ labellers. 'System pred. 2' and 'System ∞ 2' refer to predictions using the more convenient Eqs. (7) and (8) for estimating the model parameters.

7. Acknowledgements

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the project *C-AuDiT* under Grant 01IS07014B. The responsibility lies with the authors. The perception experiments were conducted by Susanne Burger (Pittsburgh) and Catherine Dickie (Edinburgh). We want to thank Andreas Maier for adapting PEAKS to our task.

8. References

[1] 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.

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

[3] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: Employing word accent information for pronunciation quality assessment of english l2 learners," in *Proc. of SLATE*, Wroxall Abbey, 2009.

[4] 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.

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

[6] 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.

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

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