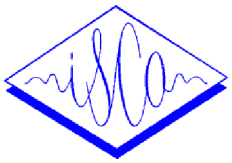
ISCA Archive
http://www.isca-speech.org/archive



Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

How Many Labellers Revisited – Naïves, Experts, and Real Experts

Florian Hönig, Anton Batliner, Elmar Nöth

Pattern Recognition Lab, Universität Erlangen-Nürnberg, Germany
{hoenig,batliner}@informatik.uni-erlangen.de

Abstract

A database of non-native German productions was annotated by three different groups: by experts using detailed, localised labels as well as coarse, global labels, and by phoneticians and naïve subjects, using the same coarse global labels. For the detailed annotation, segmental and supra-segmental labels were given segment-based and word-based. The global annotation consisted of a turn-based assessment of intelligibility, non-native accent, melody, and rhythm. Moreover, we use a large, specialised prosodic feature vector for modelling native vs. non-native speech. We study relationships between detailed and global labels, analyse the quality of expert and naïve labellers, and present an automatic system for predicting a speaker's score for the global labels.

Index Terms: non-native prosody, speech melody, rhythm, inter-labeller agreement, regression system, performance model, experts vs. naïve judges

1. Introduction

Non-native segmental and supra-segmental traits limit proficiency in a second language (L2) and by that, mutual understanding. 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. Basically, 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.

This article is a sequel of [1, 2]. In [1], we describe the global prosodic assessment of non-native English production by a large number of raters (60), and used a large prosodic feature vector and multilinear regression to predict the level of proficiency of non-native speakers of English as L2. In [2], using the same data and features, we evaluate a regression system as for its ability to predict the level of proficiency, based on 1 to n raters (labellers). We quote from the conclusion: '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.'

In the present paper, we address related but different questions: Using German as L2, we compare the performance of three different types of raters/labellers, and two different types of annotation. We employed three phonetic 'real' *experts* with extensive labelling experience, especially with the actual database, eleven *phoneticians*, i. e. students and post-grads, with no specific experience with the actual database (experts but no 'real' experts), and 18 *naïve* raters. All were native speakers of German with no known hearing loss. All three groups conducted the same global assessment experiment, cf. below. In addition, the experts annotated different aspects such as peculiarities on the segmental, word, and supra-segmental level. All annotators were paid well for taking part in the experiments, in order to ensure a high quality of the annotations.

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. In [3], it was shown that a large number of annotators ('Vox Populi') creates reliable annotations. In our experiments, phoneticians are somehow in between the two other groups of raters: They had to do the same as the naïve raters and got the same payment; chances are that they turn out to produce results in between the other two groups.

In this paper, we want to find answers to these questions: First, we assume that basically, experts are 'better' than phoneticians, and both are better than naïve raters – but does it really pay off, or can we simply compensate by employing more naïve raters than phoneticians, and more phoneticians than experts? Second, how good are we in assessing non-native traits, both by using a fine-grained, detailed expert annotation or using a global perceptual assessment? Moreover, we will describe an automatic assessment system along the lines of [1, 2].

2. Material and Human Assessment

We recorded 45 German L2 speakers: 12 French, 11 Italian, 11 Spanish (incl. 1 Catalan and 1 bilingual US English/Spanish), 5 Turkish (incl. 1 bilingual Turkish/German), 4 Russian, 1 Polish, and 1 Slovak speakers. They had to read aloud[1] 270 utterances presented by an automated recording software, and 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 consisted of the well-known story 'The North Wind and the Sun', some short dialogues, specific words, and sentences which are either phonetically balanced or illustrate specific phenomena such as sentence mood. We defined a subset that was judged by the three experienced 'real' experts as 'prosodically most error-prone for L2 speakers of German':

1. *Ich gehe auf Lothars Fete. Kommst du diesmal mit?* (I'm going

[1]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.

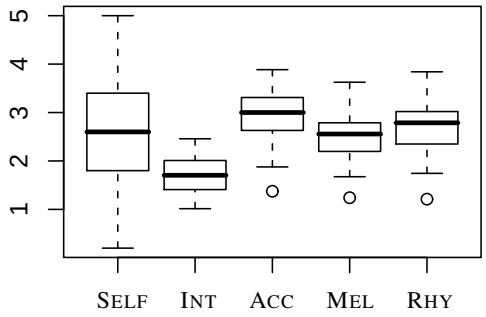


Figure 1: Box-and-whisker plot of the global annotations of phoneticians and naïves.

to Lothar's party. Will you join me this time?)

2. *Sehr gut. Ich bräuchte eine Mitgliedschaftsbestätigung. Bin ich bei Ihnen richtig?* (Very good. I do need a confirmation of membership. Is this the right place to ask for it?)
3. *Bei den Temperaturen vereisen doch die Pisten!* (At these temperatures, the piste will freeze over.)
4. *Wir hören den plätschernden Bach.* (We hear the purling creek.)
5. *Die Bremsen quietschen grässlich.* (The breakes are squeaking horribly.)
6. *Die Adidas-Aktie hinkt dem Konkurrenten Puma hinterher.* (Adidas shares are lagging behind Puma shares.)
7. *Tenor der Diskussion: Mehr Transparenz muss her!* (The tenor of the discussion: We do need more transparency!)

The global assessment was conducted as a web-based perception experiment, using the tool PEAKS [4]. The raters judged the sentences in random order. The questions were:

1. INTELLIGIBILITY (INT): DID YOU UNDERSTAND WHAT THE SPEAKER SAID?
   (1) *yes, the sentence is completely understandable* (2) *yes, but some parts are not easy to understand* (3) *well, the sentence needs some effort to be understood* (4) *no, most parts of the sentence are not easy to understand* (5) *no, the sentence cannot be understood at all*
2. FOREIGN ACCENT (ACC): DID YOU HEAR A FOREIGN, NON-GERMAN ACCENT?
   (1) *no* (2) *very slight* (3) *some accent* (4) *strong accent* (5) *extreme accent*
3. SENTENCE MELODY (MEL): THIS SENTENCE'S MELODY SOUNDS...
   (1) *normal* (2) *acceptable, but not perfectly normal* (3) *slightly unusual* (4) *unusual* (5) *very unusual*
4. RHYTHM (RHY): THE GERMAN 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*

For each speaker, the labels on the Likert scales were averaged over all seven sentences to get a single score for each criterion. Figure 1 shows a box-and-whisker plot of the resulting scores; in addition, SELF represents the speakers' self-assessment, based on a mapping of CEF [5] levels (A1 to C2) onto a corresponding scale between 1 and 5[2].

For the detailed annotation, we developed a tool for listening to arbitrary parts of the speech signal, and for annotating different tiers; for all ratings, we decided in favour of a scale consisting of 3 levels 'good', 'medium', and 'bad'. A 'pseudo-canonical' prosodic annotation and a segmental transcription were given in advance, thus the annotators only had to correct, i. e. add, delete, or substitute prosodic labels for accents and boundaries, and phonetic segments. The segmental transcription given was simply taken over from the word lexicon. The prosodic annotation 'out-of-the-blue' was of course, strictly speaking, not 'canonical' but a fair representation of a rather neutral prosody with neither too much integrating or isolating phrasal and accent structure. Here, we only detail those labels that we will use in the experiments described below:

prosody (PROS): phrase accent PA, secondary accent SA, and no accent (default, no label given); strong phrase boundary B3, and weak phrase boundary B2; each phrase delimited by B2/B2 to its right had one PA and 0-n SA. This is a fairly traditional way of representing prosody which has been described in detail in [6] – not too different from any ToBI-light version.

segmental ('phoneme') level (SEG): variants such as [r/R] or [s,z]; substitutions, deletions, and vowel/consonant insertion.

[2] A1 = 5, A2 = 4.2, ..., C2 = 1; the bilingual Turkish/German speaker was assigned a D = 0.2

Table 1: Resulting correlations when trying to predict global scores (self-assessment and averaged global scores from three experts, three phoneticians or three naïves) from the detailed annotation. 'PROS' uses prosodic annotations such as number of phrase boundaries; 'SEG' uses segmental annotations such as number of phoneme substitutions (see text).

	labeller(s)	SELF	INT	ACC	MEL	RHY
PROS	*speaker*	0.334				
	experts		0.303	0.300	0.430	0.569
	phoneticians		0.312	0.343	0.433	0.409
	naives		0.329	0.304	0.440	0.388
SEG	*speaker*	0.575				
	experts		0.919	0.823	0.810	0.811
	phoneticians		0.828	0.786	0.805	0.806
	naïves		0.842	0.815	0.799	0.778

3. Experiments and Results

3.1. Detailed vs. Global

To study relationships between detailed and global annotations, we computed *global statistics of the detailed annotations* such as the total number of consonant insertions, and compared these figures to the global annotations. In order to get the 'big picture', we tried how well we can predict the global annotations from (the global statistics of) the detailed annotations by multilinear regression. We use each detailed annotation of the three expert labellers to get an estimate of the global score, and average over the resulting three scores. As target values for the regression, we use SELF and the global annotations averaged from experts, phoneticians or naïves. The phoneticians and naïves are more numerous than the experts (11 and 18 vs. 3) which presents a problem for comparability, as labels averaged from more labellers are of higher quality and thus easier to predict. In order to compensate for this, we average the results over all permutations of three labellers from phoneticians and naïves, respectively.

Lest we overfit to the particular properties of experts or speakers, we evaluate this scheme in a nested leave-one-speaker-out and leave-one-labeller-out cross-validation[3]. We carry out this procedure separately for the detailed annotations PROS and SEG. The results are given in Table 1. One can clearly see that the PROS annotations are better at modelling

[3] For each speaker and expert labeller, the predicted global annotation is computed from a regression trained on the remaining 44 speakers and the remaining two expert labellers.

Table 2: Pair-wise labeller correlation for rhythm, within and across labeller groups. In parentheses, the correlation of the 'ground truths' between labeller groups is given.

	experts	phoneticians	naïves
experts	0.904 (1.000)	0.800 (0.971)	0.763 (0.984)
phoneticians		0.813 (1.000)	0.745 (0.997)
naïves			0.721 (1.000)

the prosodic global scores MEL and RHY than at modelling INT and ACC (e. g. 0.430/0.569 vs. 0.303/0.300 for the global scores from the experts, third row in Table 1), while the SEG annotations are better at modelling INT than at modeling the prosody-related scores MEL or RHY (e. g. 0.919 vs. 0.810/0.811 for the global scores from the experts, antepenultimate row in Table 1). However, in absolute numbers, SEG is far better than PROS in modelling any single global score. This will be discussed in Section 4.

3.2. Experts vs. Naïves

Given the task of building an automatic assessment system for target scores such as our global labels, one has to decide whether to hire experts or naïve labellers for the annotation. Thus, we now want to study the quality of the labels from the three different groups of labellers. Closely related is the question of how many to employ. Intuitively, averaging over multiple labellers will improve quality, and if one has collected annotations from a nontrivial number of labellers, one can estimate labeller and system performance for an increased number of labellers (see e. g. [2]). Along similar lines, we can make statements about different labeller groups.

Let us denote the labels from a first group of labellers $k = 1, 2, \ldots$ as X_k, and the labels from a different group of labellers $l = 1, 2, \ldots$ as Z_l. As agreement measure, we use the *Pearson correlation coefficient* between two random variables A and B, $\rho_{A,B} = \mathrm{Corr}(A, B) = \mathrm{Cov}(A, B)/\sigma_A/\sigma_B$. The annotations are modelled as jointly normally distributed random variables with $\mathrm{Var}(X_k) = \mathrm{Var}(Z_l) = \sigma^2$, $\mathrm{Cov}(X_i, X_j) = c\sigma^2 \forall i \neq j$, $\mathrm{Cov}(Z_i, Z_j) = d\sigma^2 \forall i \neq j$ and $\mathrm{Cov}(X_k, Z_l) = e\sigma^2$. That is, the pair-wise labeller correlation $\mathrm{Corr}(X_i, X_j)$ among the first labeller group is c, $\mathrm{Corr}(Z_i, Z_j)$ among the second group is d, and $\mathrm{Corr}(X_k, Z_l)$ between the groups is e.

Averaged Annotations of N labellers $X_1, X_2, \ldots, X_N$ of the first group are denoted by X^N; similarly, Z^M is the average over M labellers of the second group. From the above follows

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

We define the ground truth of the first labeller group as $L := \lim_{N\to\infty} X^N$; similarly, $K := \lim_{M\to\infty} Z^M$ is the ground truth of the second group. Thus we get $\mathrm{Corr}(X^N, K) = e/\sqrt{\frac{1}{N} + \frac{N-1}{N}c}/\sqrt{d}$ and $\mathrm{Corr}(L, K) = e/\sqrt{c \cdot d}$. The parameters c, d, and e can conveniently be estimated from given labels: Compute $\overline{\mathrm{Corr}}(X_k, X^N)$ as the average correlation of one labeller with the averaged annotation of all N labellers of the first group, and set $c := \left(N \cdot \overline{\mathrm{Corr}}(X_k, X^N)^2 - 1\right) / (N - 1)$. Similarly, d can be estimated as $d := \left(M \cdot \overline{\mathrm{Corr}}(Z_k, Z^M)^2 - 1\right) / (M - 1)$ from the average correlation of one labeller with the average of all M labellers of the second group. Then, e can be estimated using (1) from the correlation between X^N, the averaged annotation of all N labellers of the first group and Z^M, the averaged annotation of all M labellers of the second group.

Table 2 lists the estimated pairwise correlation between labellers within the groups, and across groups, for the example of the RHY score. Furthermore, the estimated correlation between the 'ground truths' of the groups is shown in parentheses. Looking at the pairwise correlation within the groups (main diagonal of Table 2), it is clear that as expected, the expert group shows the highest internal consistency ($\rho = 0.904$), followed by the phoneticians ($\rho = 0.813$), while the naïves seem relatively heterogeneous ($\rho = 0.721$). This, together with the predicted near-convergence of the ground truths ($\rho \geq 0.971$, parentheses in Table 2) explains why one expert or one phonetician correlates even higher with a naïve labeller than another naïve labeller (0.763/0.745 vs. 0.721, last column of Table 2).

Figure 2 shows the predicted behaviour of the different labels for RHY when varying the number of labellers used for averaging. Regardless of whether we aim at the ground truth of experts, phoneticians or naïves, when only one labeller is applied, 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 we set a minimum correlation of around 0.95 with any ground truth as our aim (indicated by the horizontal dashed lines in the plots in Figure 2), three experts will do the job (the right one of the two (red) squares in each plot). If we are willing to employ 5 labellers, also the phoneticians and even the naïves suffice: the intersection of the vertical dashed lines at 5 with all graphs is at around 0.95 or higher.

3.3. Automatic Assessment

Which labellers should we use for training the automatic prosody assessment system? First, of course it has to be stated that we are in an unusually and unrealistically comfortable situation with so many labellers, and taking the average of any of the three groups would work well. But now that we have the labellers available, it would be a waste not to make the most of it. Taking an agnostic attitude, we do not want to prefer any group of labellers, but rather look for an annotation that correlates highly with the ground truth of any group. We achieve this best by first computing the average annotation of each group, and then computing the average over the three groups. Similar to the computations above, we can estimate the correlation between that annotation and the ground truth of experts, phoneticians and naïves, and end up with 0.984, 0.989 and 0.993, respectively, for the example of the RHY annotations. Note that just averaging over all 3+11+18=32 labellers in an unweighted fashion would be more biased (0.979, 0.993, 0.995).

Our aim in collecting labels is to build an automatic assessment system for a learner's prosody. In order to obtain suitable input parameters for that system, we segment the recordings with forced alignment of the target utterance using a cross-word triphone HMM speech recognition system, and automatically compute a large number of features measuring different prosodic traits on speaker level. We apply our comprehensive general-purpose prosody module which has already been successfully applied to the automatic assessment of English as L2 [1]. The basic features are derived from 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 these prosodic features is given in [7]. We use these prosodic features computed over different units and contexts (e. g. the mean of the 35 features for all words, or the standard deviation of the 104 context features

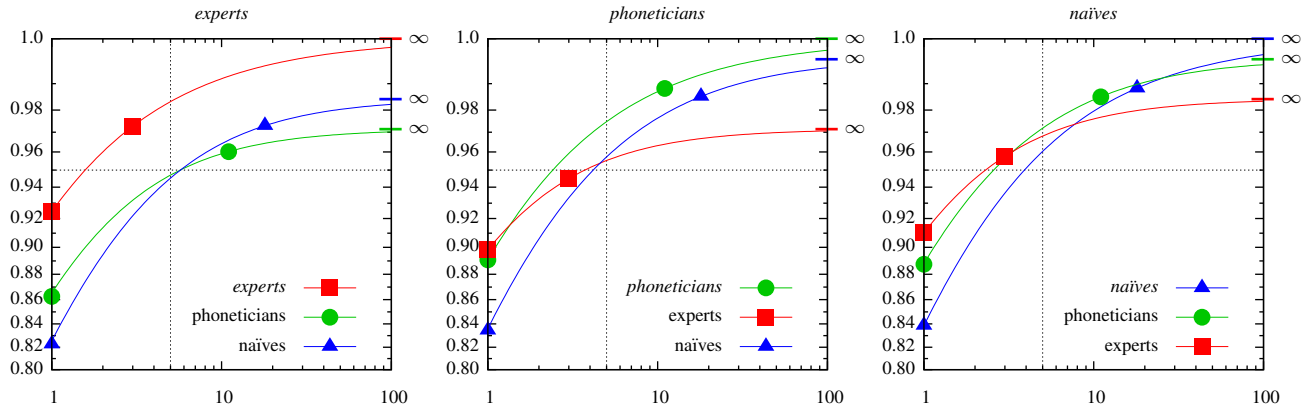


Figure 2: Predicted correlation of averaged rhythm labels from experts, phoneticians and naïves with the ground truth of *experts* (left), *phoneticians* (middle), and *naïves* (right) as a function of the number of labellers. The limits for infinitly many labellers are marked at each right y-axis with '∞'. For the labeller group that is used as ground truth in each plot, that limit is 1, e. g. the '*experts*' in the left plot. For each labeller group, the observed correlation using one labeller and using all available labellers (3 experts, 11 phoneticians, 18 naïves) is marked with a point. The x-axis is scaled logarithmically, and also the y-axis is slightly warped nonlinearly.

for all stressed nuclei) to construct generalizations of state-of-the-art rhythm features as suggested e. g. by Grabe and Low [8] and Ramus [9]. In total, we use 753 features, an extended version of the feature set described in more detail in [1].

We use a greedy forward feature selection to 5 features in a wrapper approach and ordinary multilinear regression to predict the target scores. The system's performance is evaluated in a leave-one-speaker-out cross-validation. The correlation of the system's output with the training labels[4] are: 0.539 for SELF, 0.647 for INT, 0.463 for ACC, 0.885 for RHY and 0.837 for MEL. Thus, for the example of the rhythm scores, the performance of the automatic system is better than a single naïve labeller and about as good as a single phonetician.

4. Discussion and Concluding Remarks

For a reliable annotation of specific localised phenomena, we still have to use experts; for a global assessment, naïve labellers will do as well. Employing five labellers is a good compromise [2]. For the present data, this always achieves predicted correlations of approx. 0.95 with the ground truth, regardless of whether one 'believes' in expert's or naïve's judgements. We have seen an interesting association between segmental and suprasegmental phenomena, when predicting the global assessment scores with the help of PROS and SEG: PROS does not seem to be highly predictive, even for the genuine prosodic scores MEL and RHY. The reason might be that there is much freedom in the distribution and frequency of phrase boundaries and accents – isolating and integrating speech registers are equally acceptable; however, there is a very high association between SEG and prosodic scores. Obviously, mastering segmental and prosodic traits do have much in common. Thus we can speculate that it really may pay off to automatically assess non-native speech with the help of our prosodic assessment engine, and then use these automatic scores for adjusting priors of segmental errors. Last but not least, we could demonstrate that using our general-purpose prosodic feature vector, we achieved high automatic correlations, up to .885 for RHY for our data.

[4]To estimate the system's correlation with the ground truth of one's choice, one has to multiply the given numbers by the correlation of the training labels with that ground truth, i. e. a factor around 0.98.

5. Acknowledgements

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 by Tanja Ellbogen and Susanne Waltl. We want to thank Andreas Maier for adapting PEAKS to our task.

6. References

[1] 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.

[2] 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*, 2010, no pagination.

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

[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] 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 28th June 2011.

[6] 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.

[7] 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.

[8] 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.

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