Perceiving Speech Rate Differences between Natural and Time-scale Modified Utterances

Hartmut R. Pfitzinger [1], *Hansjörg Mixdorff* [2]

[1] Pfitzinger Voice Design, Lübeck, Germany
[2] Dept. of Computer Science and Media, Beuth University, Berlin, Germany
h.pfitzinger@gmail.com, mixdorff@beuth-hochschule.de

Abstract

The effect of time-compression and -expansion on the perception of speech rate differences is investigated. Natural utterances were compared with modified versions time-scaled to the same duration. A set of ten German sentences was produced by one native speaker at slow and fast speed. In a forced choice discrimination task 15 participants were asked to select the faster one of two versions of the same sentence. In the case of low speech rate, versions that had been slowed down were perceived as slower than the corresponding natural utterances, whereas at high speech rates, stimuli with increased speed were judged as relatively faster. The effect turned out to be stronger for the slow stimuli. These findings suggest that the underlying articulatory effort plays an important role in the perception of speech rate.

Index Terms: speech rate, time-scale modification, discrimination test, perception, articulatory effort

1. Introduction

It is a well-known fact that articulatory movements are more reduced at higher speech rates. However, the relationship between articulatory effort [1] and speech rate is not trivial. Kuehn&Moll [2], for instance, found that speakers use individual strategies of decreased tongue displacements and increased tongue velocity to produce a higher speech rate. Tillmann&Pfitzinger [3] showed via electro-magnetic mid-sagittal articulography (EMMA) on the sentence level that displacements of the articulators are not only reduced but also that complex trajectories are simplified and even skipped at very high speech rates. These simplifications of articulatory movements lead to less articulatory effort compared with the canonical forms. On the other hand, slow speech rates lead to utterances produced with more precision and thus with higher articulatory effort (see also Hyper&Hypo theory [4]).

These findings raise the question whether two utterances by the same speaker produced with identical duration, but different articulatory precision lead to different perceived speech rates. In other words, do listeners take into account the articulatory effort when judging speech rate?

This problem is of special importance in the context of the different dimensions of prosody as postulated by Firth 1948 [5]. The three dimensions Intensity, Intonation, and Timing have been widely investigated and several approaches to their formal description exist, whereas Voice Quality, termed the 4th dimension by Campbell&Mokhtari [6], and its suprasegmental variations have proved difficult to formalize [7]. In addition, the degree of reduction, termed the 5th dimension of prosody by the first author [8], was touched on in early studies [9, 10], but to-date has not been described by any model, for good reasons. Not only does it require an immense amount of phonological, but also segmental knowledge since e.g. the estimation of undershoot and overshoot would probably be a component of a model of the prosody of the degree of reduction. "Articulatory effort" appears hard to quantify and is closely related to the seemingly concrete "degree of reduction".

While the degree of reduction, reinterpreted as a prosodic dimension, hitherto escaped acoustic measurement there are numerous studies and methods for measuring speech rate.

Ohno&Fujisaki [11] developed a method for calculating *relative local speech rate*, that is, the local ratio between the temporal structures of two versions of the same sentence. For this purpose, a function is determined via *dynamic time warping* (DTW) which optimally maps the time axis of one utterance to that of another one. If a passage of one signal is shorter than the corresponding one of the other signal, the warping function produces a local expansion. Smoothing the warping function by a window of 300ms yields the relative local speech rate.

This purely acoustic definition would yield a mean value of 1 between two utterances with the exact same durations and therefore identical speech rates, even though the articulatory efforts and segmental structures could differ considerably.

Pfitzinger investigated speech rate in a number of perceptual experiments and developed a model for predicting auditory judgements based on phone and syllable segmentations of the underlying acoustic signal with sufficient accuracy [12, 13]. This model of *perceptual local speech rate (PLSR)* has been applied successfully in a number of prosodic studies (e.g. [14, 15, 16]). Articulatory precision, however, is treated rather coarsely, namely, when it manifests on the segmental level as speech segments being added (insertions, e.g. epenthetic vowels) or deleted completely (elisions).

The prediction of PLSR, however, leaves a variance of up to 19% unexplained [13] and even the consideration of fundamental frequency which has been shown to influence speech rate perception [17] and therefore indicates significant interactions between the two prosodic dimensions, did not significantly resolve this discrepancy. This suggests that part of the unexplained variance could be attributed to the degree of reduction, since insertions and deletions have a clear influence on the predicted PLSR values. In that case, an additional interaction could be quantified, namely that between PSLR and the degree of reduction.

Quantifying the degree of reduction, however, is very difficult. Therefore we shall apply manipulations to the speech signal in order to approximate the solution: if the assumption formulated above is correct, the perceived degree of reduction could be estimated by comparing perceptual speech rate judg-

Table 1: The ten sentences used in the perception experiment and their English translations.

	German sentence	English translation
1.	Male nie nur Rosen.	Never paint roses only.
2.	Plötzlich strickst du flinker.	Suddenly you knit faster.
3.	Immer bei Tagesanbruch gehe ich durchs Moor zur Arbeit.	Always at dawn I go through the moore to work.
4.	Das ist der schnellste Weg zu den Kaligruben hinter den Hügeln.	This is the fastest way to the potash pits behind the hills.
5.	Eine Thermoskanne voll Kaffee ist im Rucksack verstaut.	A thermos full of coffee is stowed in the backpack.
6.	Außerdem ein Riegel Schokolade, eine Banane und mein Mittagessen.	Besides, a bar of chocolate, a banana and my lunch.
7.	Ich esse meist dort, wo ich arbeite, verlasse die Grube nicht.	I eat mostly (there) where I work, do not leave the pit.
8.	Die anderen gehen in die mobile Kantine am Rand der Kaliwerke.	The others go to the mobile canteen at the edge of the potash works.
9.	Man hat sie extra für uns in Kanada gekauft, heißt es jedenfalls.	It was bought specially for us in Canada, at least that's what is said.
10.	Ich soll doch ein wenig geselliger sein, sagt Lisa, meine Frau.	I should just be a bit more sociable, says Lisa, my wife.

ments on utterances of the same durations, but different degrees of reduction. Therefore we create stimuli for a perceptual experiment from utterances of the same sentence produced at different speech rates. In one case, fast versions will be decelerated to match the durations of normal, slowly spoken versions, in the other, slower versions will be accelerated to yield the durations of naturally uttered fast versions.

To this end, a reliable, artifact-free technique is required by means of which the duration of the signal can be modified without affecting its spectral properties. Janse [18, p.16–19] discussed the perception of technically compressed or expanded speech in detail. In particular, she examined the method for manipulating the time axis of recorded utterances developed by Covell, Withgott&Slaney [19] which takes phonetic knowledge into account by treating pauses, stressed vowels, unstressed vowels, and consonants in a non-uniform fashion.

Janse [20] showed that moderately time-manipulated speech is perceived as natural and that different algorithms for compressing or expanding the signal do not significantly affect intelligibility. Therefore, the approach by Covell et al. [19] did not yield any particular gain over other methods. Janse also showed that linearly time-compressed utterances produced at normal speech achieved higher word recognition rates than regular fast uttered versions.

1.1. Hypothesis

We expect that the speech rate of natural fast utterances will appear lower than that of hyperarticulated utterances of the same duration, produced by time-compressing slow utterances. In other words, the smaller amount of reduction in the hyperarticulated utterances will trigger the perception of higher articulatory effort and therefore a higher perceptual speech rate.

In contrast, technically decelerated utterances should appear slower than the naturally produced utterance with the same duration, since they reflect a smaller articulatory effort.

Provided these utterances are judged significantly different, it would indicate that the precision of phonetic realization has a direct impact on the perceptual speech rate, beyond the effects of segment insertions and deletions. Such a result would also confirm that speech rate is not determined by the canonical form underlying an utterance, but by its actual realization.

In summary, the experiment presented in the following sections was performed to dismiss one of these hypotheses:

H_0 = "Different phonetic realizations of the same sentence are perceived as equally fast as long as their durations are equal."

H_A = "Different phonetic realizations of the same sentence and equal duration are consistently perceived as different with respect to their speech rate."

2. Method

2.1. Stimulus selection and recording

The underlying sentences are listed in Table 1 and were selected as they had been employed in other studies on speech rate and were available in several versions.

A male phonetician (age 35) produced the ten sencences in the three speech rates *normal*, *fast*, and *slow*. They were recorded in an anechoic chamber with the high-quality close-talk microphone (*Beyerdynamic NEM 192*), digitized at 48 kHz and 16 Bit and downsampled to 16 kHz.

2.2. Stimulus manipulation

We employed our own implementation of TD-PSOLA [21] for expanding and compressing the speech data which differs from Covell et al. [19] in that the time-scale is manipulated linearly.

All three natural versions (*normal*, *fast*, and *slow*) of the 10 sentences were manipulated. Hence we created four artificial variants: the fast version adjusted to the slow one (decelerated to 56%), the normal version adjusted to the fast one (accelerated to 140%) and transformed to match the slow one (decelerated to 71%), and the slow version transformed to the fast one (accelerated to 180%). In addition to the four artificial versions the fast and slow natural utterances were employed. The stimuli at mid speed were only created to derive the slightly accelerated and decelerated variants.

Figure 1 shows the sentence "Male nie nur Rosen" (engl.: "Never paint roses only") produced at a slow speech rate, and time-expanded to 140% of the original duration and 180%, respectively.

2.3. Stimulus randomization

18 combinations for each of the ten sentences yielded 180 stimulus pairs. The 18 combinations were created from the six versions of each sentence as follows: all combinations of the three "fast" stimuli (natural, moderately accelerated, strongly accelerated) and all combinations of the three "slow" stimuli (natural, moderately decelerated, strongly decelerated). Hence each comparison appears twice, though in reverse order, except for the pairs of identical stimuli.

The resulting 180 stimuli were randomized and inserted between five dummy pairs at the beginning and five at the end, respectively, yielding a total of 190 stimuli.

2.4. Presentation to the listeners

15 subjects participated in a clocked listening test in which the stimuli were divided into groups of ten preceded by a long beep and a short beep before each stimulus pair with a pause of three

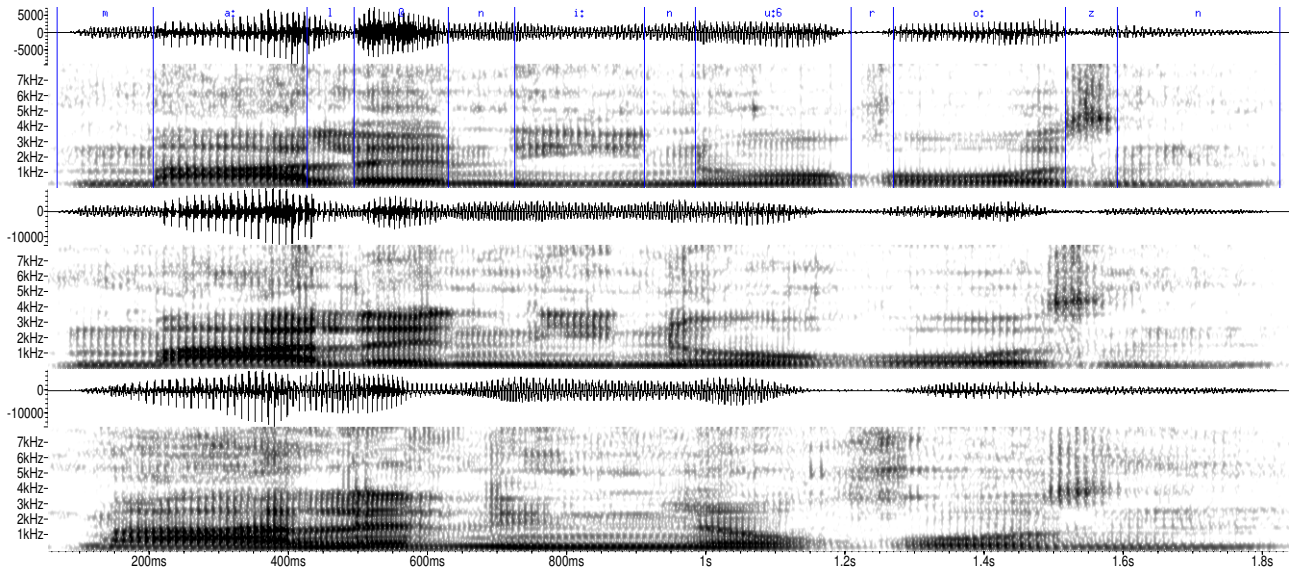

Figure 1: Oscillograms and sonagrams of the German sentence "Male nie nur Rosen" (engl.: "Never paint roses only") [ma:lə ni: nuʁ ʁo:zn]. *Top:* slow speech rate, *middle:* normal speech rate time-expanded to 140%, *bottom:* fast speech rate time-expanded to 180%.

seconds between the tasks. Each stimulus pair was played back once and a forced decision had to be made which of the two utterances had been perceived as faster. The stimuli were presented via high-quality closed headphones (*Beyerdynamic DT-770Pro*).

3. Results

The judgments from the discrimination task were split into three groups: the first group consisted of 60 stimulus pairs with identical stimuli, the second group contained 60 pairs at low speed, and the third one 60 pairs at high speed.

Results from the first group with the pairs of identical stimuli revealed that listeners showed individual biases towards the first or second stimulus in a pair. In order to avoid this bias at the evaluation of the second and third group, instead of adding and averaging the judgments, we calculated for each listener and sentence the difference between the two judgments on pairs of identical stimuli in the normal and the reverse order. These differences were added over all listeners, yielding a maximum value of 15, if all listeners decided in favour of the first stimulus and a minimum of -15 in the opposite case. These extremes were mapped to $\pm 100\%$. A value of 0% therefore means that none of the stimuli in a pair appeared to be faster.

Figure 2 shows these results for all slow utterances, the average over all 10 sentences being represented by one box plot. The box shows the lower quartile, median, and upper quartile values. The whiskers indicate the extent of the remaining data. Outliers are marked with a cross. The notches represent a robust estimate of the uncertainty about the means for box-to-box comparison. Table 2 shows the results of ANOVA of the slow stimuli. Results for the group of fast stimuli are displayed in Figure 3 and Table 3.

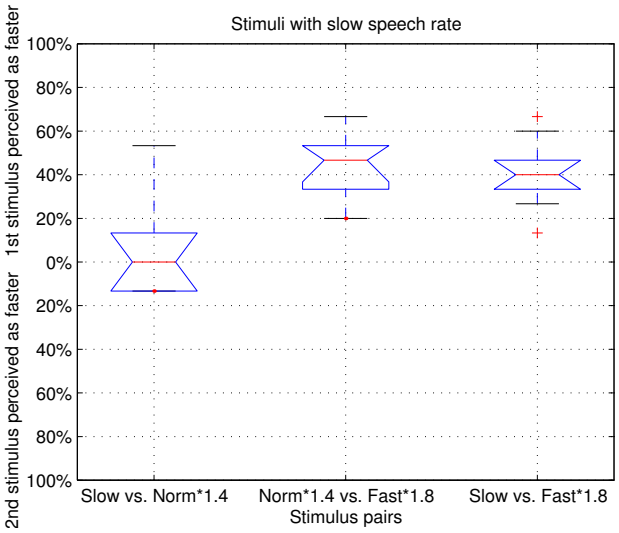

Figure 2: Perception results for the slow stimuli.

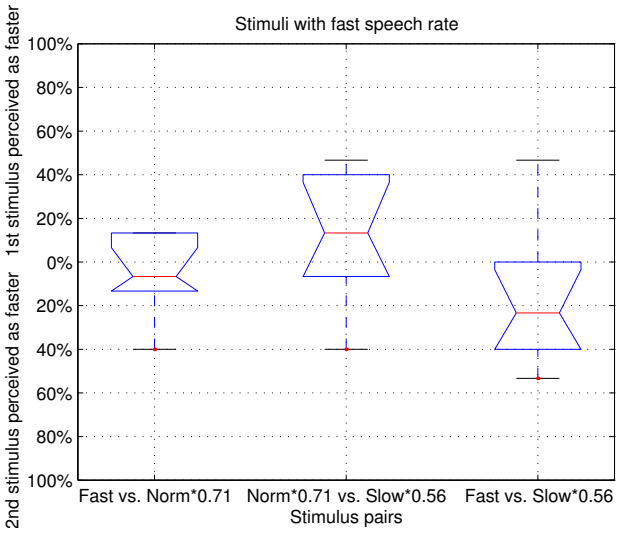

Figure 3: Perception results for the fast stimuli.

Table 2: ANOVA of the listeners' judgements of the slow stimuli with the factors *stimulus pair* and *sentence*.

	df	F	p	expl. variance
Stimulus pair	2	14.667	0.0002	46.3%
Sentence	9	1.787	0.1408	25.4%

Table 3: ANOVA of the listeners' judgements of the fast stimuli with the factors *stimulus pair* and *sentence*.

	df	F	p	expl. variance
Stimulus pair	2	6.907	0.0059	16.7%
Sentence	9	5.679	0.0009	61.6%

4. Discussion

The speech rates inside the groups of fast and slow stimuli were judged significantly different ($p<0.001$, Table 2 and $p<0.01$, Table 3, respectively). Thus, H_0 must be rejected (see Section 1.1). Hence it needs to be stated that neither the linguistic content nor the underlying canonical form determined the observed perceptual speech rate differences. In the opposite case, all utterances of the same duration would have been been judged as equally fast.

Therefore, the model of *relative local speech rate* [11] is not suitable for predicting the speech rate differences perceived by the listeners, since it yields the same relationship between utterances of equal duration, namely unity, regardless of the degree of reduction at which they were produced. The model of PLSR has the advantage of considering insertions and deletions, however, is incapable of predicting the subtle perceptual differences observed in the current experiment.

We found considerable asymmetries regarding the strength, the uniformity, and the direction of these differences. Pair-wise comparison reveals that in the case of slow stimuli the originally fast but strongly decelerated utterances are always perceived as slower (Figure 2), whereas in the case of fast stimuli only the strongly accelerated utterances compared with the moderately accelerated ones are perceived as faster.

The results for the fast stimuli are not as pronounced as those for the slow stimuli. Individual analysis of the ten sentences yielded the following observation which, however, needs to be verified in follow-up studies: pauses, which occasionally appear in the slowly spoken utterances and do not disappear after acceleration seem to reduce the perceived speech rate, even if the remaining segments are compressed to produce the same utterance duration.

A serious problem of our method is the fact that some time-scale modifications violate articulatory constrains. However, since all signals are compressed or expanded by less than factor two, according to Janse [20] even our "strong" manipulations are still moderate and produce acceptable stimuli. At worst, listeners commented that strongly decelerated stimuli sounded as if spoken by a drunk person.

Janse [20] showed that normal speech that was technically accelerated yielded shorter reaction times on target word recognition than fast speech. She drew the conclusion that technically accelerated speech was more intellegible than natural fast speech. Based on the results of the current perception study we can now posit that speech accelerated by more than 40% sounds faster than natural speech at the same syllable rate.

We conclude therefore that the cause of this difference lies in the smaller degree of reduction and the higher articulatory precision of the original slow utterance connected with higher articulatory effort. When such an utterance is accelerated considerably, the perceived articulatory effort will increase accordingly: due to the higher density of acoustic cues and probably increased redundancy the utterance will appear faster, though it exhibits the same duration as the natural fast utterance.

5. Conclusions

All fast realizations of a sentence used in the present perception experiment contain the same words, the same number of syllables and even the same number of phonetic segments, irrespective of the varying degree of reduction caused by the two different amounts of time-compression. Nevertheless, they are perceived as having significantly different speech rates. The same is valid for the slow utterances. These results suggest that the underlying articulatory effort plays an important role in the perception of speech rate.

The present experimental design does not yield quantitative results regarding the relationship between the degree of reduction and the perceptual speech rate.

In future perception experiments an adjustment task will be used to determine the utterance durations which equal the perceptual speech rate differences found in the current study. This way we achieve a quantitative representation of the perceptual behaviour. Furthermore, development of acoustic measures which relate to the degree of reduction, e.g. MFCC-deltas averaged over an utterance or the dynamic vowel quality trajectories [22, 23, 24], is indispensable for refining any speech rate model and particularly the model of PLSR to predict perceptual local speech rate.

6. References

[1] W. L. Nelson, "Physical principles for economies of skilled movements," *Biological Cybernetics*, vol. 46, pp. 135–147, 1983.

[2] D. P. Kuehn and K. L. Moll, "A cineradiographic study of VC and CV articulatory velocities," *J. of Phonetics*, vol. 4, pp. 303–320, 1976.

[3] H. G. Tillmann and H. R. Pfitzinger, "Local speech rate: Relationships between articulation and speech acoustics," in *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 3, Barcelona, 2003, pp. 3177–3180.

[4] B. E. F. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*, ser. Nato ASI series D: Behavioural and social sciences, W. J. Hardcastle and A. Marchal, Eds. Dordrecht, Boston, London: Kluwer Academic Publishers, 1990, no. 55, pp. 403–439.

[5] J. R. Firth, "Sounds and prosodies," *Transactions of the Philological Society*, pp. 127–152, 1948.

[6] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 3, Barcelona, 2003, pp. 2417–2420.

[7] H. R. Pfitzinger, "Segmental effects on the prosody of voice quality," *J. of the Acoustical Society of America*, vol. 123, no. 5, p. 3424, 2008.

[8] ——, "Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction," in *Speech Prosody Abstract Book. Studientexte zur Sprachkommunikation*, R. Hoffmann and H. Mixdorff, Eds. Dresden: TUDpress, 2006, vol. 40, pp. 6–9.

[9] P. Menzerath and J. M. de Oleza S. J., *Spanische Lautdauer. Eine experimentelle Untersuchung*. Berlin, Leipzig: Walter de Gruyter, 1928.

[10] H. G. Tillmann, "Silbischer Ausprägungskode und Intonation," *Acta Universitatis Carolinae: Philologica 1, Phonetica Pragensia*, vol. III, pp. 261–265, 1972.

[11] S. Ohno and H. Fujisaki, "A method for quantitative analysis of the local speech rate," in *Proc. of EUROSPEECH '95*, vol. 1, Madrid, 1995, pp. 421–424.

[12] H. R. Pfitzinger, "Local speech rate as a combination of syllable and phone rate," in *Proc. of ICSLP '98*, vol. 3, Sydney, 1998, pp. 1087–1090.

[13] ——, "Phonetische Analyse der Sprechgeschwindigkeit," Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, Forschungsberichte (FIPKM) 38, 2001.

[14] H. Mixdorff and H. R. Pfitzinger, "Analysing fundamental frequency contours and speech rate in Map Task dialogs," *Speech Communication*, vol. 46, no. 3–4, pp. 310–325, 2005.

[15] H. R. Pfitzinger and M. Tamashima, "Comparing perceptual local speech rate of German and Japanese speech," in *Proc. of the 3rd Int. Conf. on Speech Prosody*, vol. 1, Dresden, 2006, pp. 105–108.

[16] N. Amir, H. Mixdorff, O. Amir, D. Rochman, G. M. Diamond, H. R. Pfitzinger, T. Levi-Isserlish, and S. Abramson, "Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting," in *Proc. of the 5th Int. Conf. on Speech Prosody*, Chicago, 2010.

[17] K. J. Kohler, "Parameters of speech rate perception in German words and sentences: Duration, F_0 movement, and F_0 level," *Language & Speech*, vol. 29, pp. 115–139, 1986.

[18] E. Janse, "Production and perception of fast speech," Ph.D. dissertation, Institute of Linguistics OTS, Utrecht; Niederlande, 2003.

[19] M. Covell, M. Withgott, and M. Slaney, "Mach1: Nonuniform time-scale modification of speech," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP98)*, vol. 1, Seattle, 1998, pp. 349–352.

[20] E. Janse, "Intelligibility of time-compressed speech: Three ways of time-compression," in *Proc. of ICSLP 2000*, vol. 3, Beijing, 2000, pp. 786–789.

[21] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5/6, pp. 453–467, 1990.

[22] H. R. Pfitzinger, "Dynamic vowel quality: A new determination formalism based on perceptual experiments," in *Proc. of EUROSPEECH '95*, vol. 1, Madrid, 1995, pp. 417–420.

[23] ——, "Acoustic correlates of the IPA vowel diagram," in *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 2, Barcelona, 2003, pp. 1441–1444.

[24] ——, "Towards functional modelling of relationships between the acoustics and perception of vowels," in *Speech Production and Perception: Experimental Analyses and Models. ZAS Papers in Linguistics*, S. Fuchs, P. Perrier, and B. Pompino-Marschall, Eds. Berlin: Zentrum für Sprachverarbeitung, 2005, vol. 40, pp. 133–144.