Frequency distributions of uniphones, diphones, and triphones in spontaneous speech

Victor Kuperman
Radboud University Nijmegen, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

Mirjam Ernestus
Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands and Radboud University Nijmegen, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

Harald Baayen
University of Alberta, 4-32 Assiniboia Hall, Edmonton, Alberta T6G 2N8, Canada

(Received 16 August 2007; revised 2 October 2008; accepted 3 October 2008)

This paper explores the relationship between the acoustic duration of phonemic sequences and their frequencies of occurrence. The data were obtained from large (sub)corpora of spontaneous speech in Dutch, English, German, and Italian. Acoustic duration of an n-phone is shown to codetermine the n-phone's frequency of use, such that languages preferentially use diphones and triphones that are neither very long nor very short. The observed distributions are well approximated by a theoretical function that quantifies the concurrent action of the self-regulatory processes of minimization of articulatory effort and minimization of perception effort.
 [DOI: 10.1121/1.3006378]

PACS number(s): 43.70.Bk, 43.70.Fq, 43.70.Mn, 43.70.Kv [AL] Pages: 3897–3908

I. INTRODUCTION

Speech inherently unfolds in time and the acoustic duration of speech units is one of the characteristics of speech that is directly experienced by both speakers and listeners (e.g., Ohala, 1996). Research of the past decades has established a large variety of phonological and prosodic factors affecting acoustic duration of n-phones and syllables. For instance, stressed syllables are realized longer than unstressed ones (e.g., Ladefoged, 1982) and words at the beginning and the end of utterances show articulatory strengthening (e.g., Bell *et al.*, 2003; Cambier-Langeveld, 2000; Fougeron and Keating, 1997). Furthermore, phonemes are realized shorter the greater the number of syllables or segments in the word (Nooteboom, 1972).

In addition, the variability in acoustic duration is codetermined by the predictability of a speech unit given its phonological, lexical, semantic, syntactic and discourse contexts (e.g., Bard *et al.*, 2000; Bolinger, 1963; Fowler and Housum, 1987; Jurafsky *et al.*, 2001; Lieberman, 1963). The more predictable a phoneme, morpheme, syllable, or a word is in its context, the less important the acoustic signal is for recognition of such a unit, and the shorter it is realized (e.g., Aylett and Turk, 2004, 2006; Van Son and Van Santen, 2005). For example, function words are more likely to be realized longer when they are unexpected, i.e., less predictable in the sentence (Bell *et al.*, 2003). Similarly, phonemes that are important for word disambiguation and thus are less predictable from the preceding phonemes are less reduced, as indicated among others by their longer acoustic duration (Van Son and Pols, 2003).

Starting with Zipf (1929; 1935), the frequency of occurrence of a speech unit has been considered as an important codeterminer of its predictability and has been argued to enter into a negative (linear or nonlinear) relation with the degree of articulatory complexity of that unit (cf. Pluymaekers *et al.*, 2005). Since we consider acoustic duration as an approximation of articulatory complexity (see discussion below), Zipf's (1935) approach can be reinterpreted such that the frequency of a unit predicts its duration. Hence we label this approach "frequency predicts duration" (FPD).

The present paper explores an alternative view of the relationship between acoustic duration and frequency of occurrence such that we consider frequency of use as a function of acoustic duration and not vice versa. The advantages of this approach, which we label "duration predicts frequency" (DPF), will be pointed out in the body of the paper.

The objects of this study are uniphones, and also larger sequences of phones, i.e., diphones and triphones. Since articulatory gestures typically stretch over the boundaries of individual phones, larger phone sequences are more stable units than uniphones and we may obtain more reliable results for these longer speech units. Similar considerations have led to the common use of diphones (or larger blocks of speech) as basic units in automatic speech recognition (e.g., Richardson *et al.*, 2003) and speech synthesis (e.g., O'Shaughnessy *et al.*, 1988). We study n-phones in spontaneous speech, as it is a more natural speech variety than, say, careful speech or the speech production conditioned by experimental tasks.

We begin with reporting the consistent functional relationship between n-phone frequency and duration and show that our approach yields a better approximation to empirical data than Zipfian FPD models. Since acoustic duration is in itself influenced by multiple factors, we then confirm that this relationship also holds when effects of these predictive factors are partialled out from our estimates of acoustic duration.

TABLE I. (Sub)corpora used for data collection.

Language	Corpus	Subcorpus	No. of phonemes	Hours	Speakers
Dutch	IFA	Spontaneous monologues	36 000	1	8
American English	Buckeye	Dialogues	431 000	22	20
German	BAS	German-German dialogues	1 976 000	54	1139
Italian	AVIP	Dialogues between adults	28 000	0.6	22

We studied n-phone frequencies in Dutch, English, German, and Italian. The primary reason for selecting these languages was the availability of large (sub)corpora of spontaneous speech for those languages. Also, the languages represent two language families, Germanic and Romance, which allow for generalizability of the results. We note that even though three of the languages we consider are Germanic, they vary in the size of their phonemic inventories (and thus in frequencies of individual phones), as well as in their phonologies (e.g., final devoicing in German and Dutch, but not in English, which affects uniphone frequencies in these languages), as well as in their affixes and the frequencies of these affixes, which affect the frequencies of the n-phones (e.g., Baayen, 1994).

In order to obtain a better understanding of the observed cross-linguistic patterns, we model the relation between frequency and acoustic duration of n-phones. We fit our data with a model based on the interaction of the speaker's tendency to minimize articulatory effort (e.g., produce less clear speech) and the listener's tendency to minimize perception effort (e.g., prefer clearer speech)(Job and Altmann, 1985).

II. METHODOLOGY

A. Corpora of spontaneous speech

The data for this study were obtained from four corpora with extensive collections of spontaneous speech: The IFA spoken language corpus of Dutch (IFA) (Van Son *et al.*, 2001), the Buckeye speech corpus for American English, version 1 (Buckeye) (Pitt *et al.*, 2005), modules Verbmobil-I and -II of the Bavarian speech archive for German (BAS) (Schiel *et al.*, 1997), and the spoken Italian varieties archive for Italian (AVIP) (Scuola Normale Superiore di Pisa, 2001); see Table I for descriptions of these spontaneous speech (sub)corpora. In these corpora, speakers were not forced to use a very high or a very low speech rate, so we restrict our findings to a "normal" self-paced range of speech rates.

The speech files of these corpora come with transcriptions at the phone level. Moreover, these transcriptions provide temporal boundaries for each phone in the signal (i.e., phone-level aligned segmentation). Except for the manually aligned IFA corpus, all collections were labeled automatically with subsequent manual verification of the alignment.

Our investigations assumed the segment inventories for the four languages that formed the basis for the labeling conventions used in the respective corpora. The only exception was that we reclassified nasalized vowels in American English as oral vowels. This adjustment affected less than 0.5% of the total number of phones in the Buckeye corpus.

B. Variables

For each language, we calculated the frequency of occurrence of every uniphone in the respective corpus. This measure, *frequency*, was considered as the dependent variable. The main predictor of interest to us, *duration*, was estimated for each dataset as the average duration of the uniphone. The type of uniphone, vowel or consonant (*type*), served as a control variable. Each language was fitted with a separate multiple regression model. We then extended our survey to diphones and triphones, fitting one statistical model to the diphones and one statistical model to the triphones in every language.

We defined diphones (or triphones) as sequences of two (or three) phones without an intervening pause, end of turn, noise, laughter, a nonspeech sound, a phone marked as incomprehensible by the transcribers, or a segment extraneous to the segment inventory of that language. Notably, in identifying the diphone or triphone sequences, we ignored word or utterance boundaries. That is, we started from the first diphone or triphone and moved through the whole corpus shifting the sampling window one phone at a time. Thus, the English word "cow" [kaʊ] in a corpus would give rise to three uniphones ([k], [a], and [ʊ]), two diphones ([ka] and [aʊ]) and one triphone ([kaʊ]). This approach treats the speech signal as a continuous stream, in which word segmentation is not a given, but rather a task for the listener (e.g., Cutler and Clifton, 1999).

For the diphones and triphones, again, frequency was the dependent variable, while the mean duration of the sequences, duration, was the key predictor. We also coded the segments in the diphones as C (for consonant) or V (for vowel), which gave rise to four levels: CC, CV, VC, and VV. The control variable type for triphones had eight levels.

III. RESULTS

In all analyses reported below, frequencies of occurrence as well as durations were (natural) log transformed in order to remove most of the skewness from the distributions. The logged durations were subsequently normalized by subtracting the minimum value of duration and dividing the difference by the maximum acoustic duration in the dataset: As a result, acoustic durations ranged from 0 to 1.

A. Uniphones

For each of the four datasets with uniphones, we fitted a stepwise multiple regression model with frequency as the dependent variable. Data points that fell outside the range of −2.0 to 2.0 units of standard deviation (SD) of duration or of

TABLE II. Models of uniphone, diphone, and triphone frequencies. In column "Count," the first figure shows the total number of data points, while the figures in parentheses show the numbers of data points remaining in the model after removal of outliers. Column "Duration" lists the regression slopes for uniphone durations, and the slopes for the first and the second coefficients of the restricted cubic spline for durations of diphones and triphones, while the next column shows their p-values. Column "Type" presents the F-values for type and the next column shows their p-values. Column ΔR^2 shows the unique contribution of duration to the explained variance of the model.

Language	Count	Duration ($\hat{\beta}$)	p	Type (F-value)	p	R^2	Residual st. error	D.f.	ΔR^2
				Uniphones					
Dutch	37(33)	−0.62	0.37	0.02	0.89	0.0	0.92	30	0.0
English	45(39)	−2.01	<0.0001	18.56	<0.0001	0.41	0.59	36	0.38
German	40(37)	−0.28	0.65	3.72	0.06	0.06	0.88	34	0.0
Italian	71(66)	−1.67	0.01	0.81	0.37	0.11	1.78	63	0.09
				Diphones					
Dutch	1002(937)	First: 1.34	0.004	47.72	<0.0001	0.15	1.22	931	0.03
		Second: −2.71	<0.0001						
English	1855(1763)	First: 0.37	0.38	112.99	<0.0001	0.19	1.57	1757	0.07
		Second: −3.33	<0.0001						
German	1390(1299)	First: 4.54	<0.0001	55.77	<0.0001	0.20	2.17	1293	0.06
		Second: −8.37	<0.0001						
Italian	939(851)	First: 1.45	0.002	16.87	<0.0001	0.09	1.25	845	0.05
		Second: −3.08	<0.0001						
				Triphones					
Dutch	6909(6212)	First: 0.53	<0.0001	47.46	<0.0001	0.06	0.57	6202	0.01
		Second: −0.89	<0.0001						
English	29804(26826)	First: 1.16	<0.0001	217.6	<0.0001	0.09	0.87	26816	0.04
		Second: −2.01	<0.0001						
German	18854(16944)	First: 3.10	<0.0001	76.62	<0.0001	0.08	1.48	16934	0.05
		Second: −4.81	<0.0001						
Italian	4425(4038)	First: 0.89	<0.0001	24.88	<0.0001	0.07	0.73	4028	0.03
		Second: −1.58	<0.0001						

frequency were excluded from the analysis prior to fitting the models. After the initial fit, data points that had Cook's distance (a measure of the effect of deleting a data point) exceeding 0.2 were removed and the models were refitted.

Table II (uniphones) summarizes the findings for the uniphones in the four datasets. In the second column of this table, the first number shows the total number of data points, while the number in parentheses shows the number of data points after removal of all outliers. The third and fourth columns present the regression coefficients and p-values for duration and the fifth and sixth presents F-values and p-values for type, respectively. The last column in the table shows the unique contribution of duration to the explained variance of the model.

The predictivity of acoustic duration for the frequency of the uniphones' occurrences differs across languages. Where such predictivity is statistically significant (English and Italian), our models replicate the findings by Zipf (1935): The articulatory complexity of a phoneme (approximated here as a phoneme's acoustic duration) is inversely related to its frequency of occurrence. That only two out of the four languages demonstrate a significant correlation may relate to the fact that the duration of a segment is codetermined by the quality of its neighboring segments due to coarticulation. We may therefore expect the diphones and triphones to show more consistent correlations across languages.

Throughout this paper we used a restricted cubic spline with 3 knots (see, e.g., Harrell, 2001) to estimate nonlinear relationships between duration and frequency. For the uniphones, we found none. Moreover, none of the models for uniphones showed significant interactions between duration and type.

B. Diphones

Multiple regression models were then fitted to the four datasets of diphones. Data points that fell outside the range of −2.0 to 2.0 units of SD of duration or of frequency were again excluded from the analysis. For all data points, Cook's distance was less than 0.2. Table II (diphones) reports the results of this model fitting.

The main variable of interest, duration, was a significant nonlinear predictor of diphone frequency across all datasets. In addition, type was significant. None of the models showed significant interactions between these two predictors. Figures 1(a)–1(d) show the distributions of the frequencies of the diphones over their durations in the four languages with addition of the polynomial regression lowess smoother lines (Cleveland, 1979).

Importantly, we find that in all datasets with diphones (and in all regression models) the functional relation between duration and frequency shows concave curves, rather than

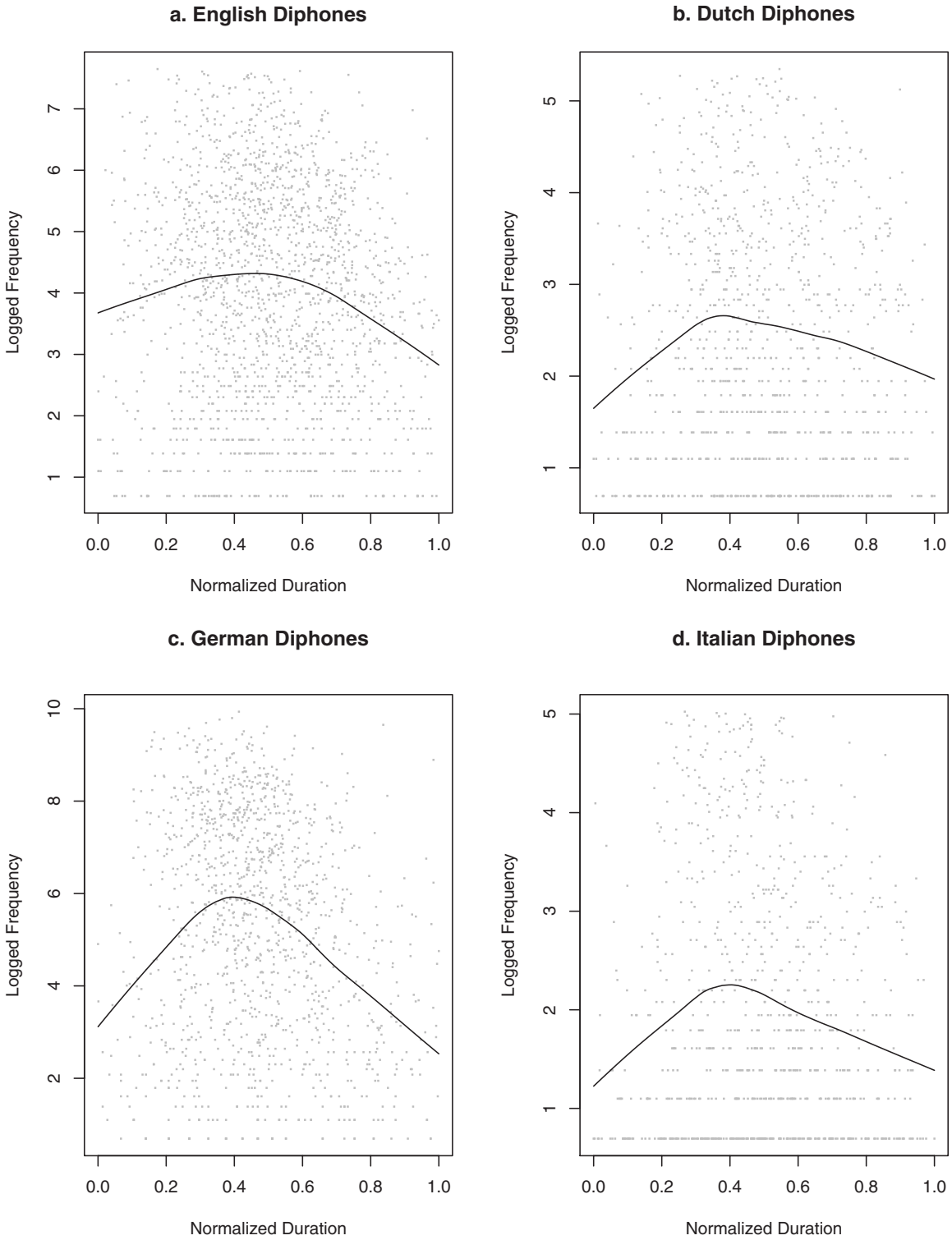

FIG. 1. Distribution of the diphone frequencies over their acoustic durations.

the monotonically decreasing curves predicted by Zipf's (1935) approach. The maxima of the curves are asymmetrically shifted leftwards toward the shorter durations, in all languages. In general, long and very short diphones are less frequently used in the four languages than diphones from the short-to-mid range of the durational spectrum.

The fact that the shortest diphones are not of a high frequency hints at the sensitivity of speakers to the discriminability of the speech signal: The shorter the duration, the more effort is required for speech perception. At the same time, long diphones are disfavored, possibly since they may take more effort to produce. We will return to this issue below.

C. Triphones

Finally, we modeled for each of the four languages triphone frequency as a function of acoustic duration and CV type. Data points that fell outside the range of −2.0 to 2.0 units of SD of duration or of frequency were excluded from

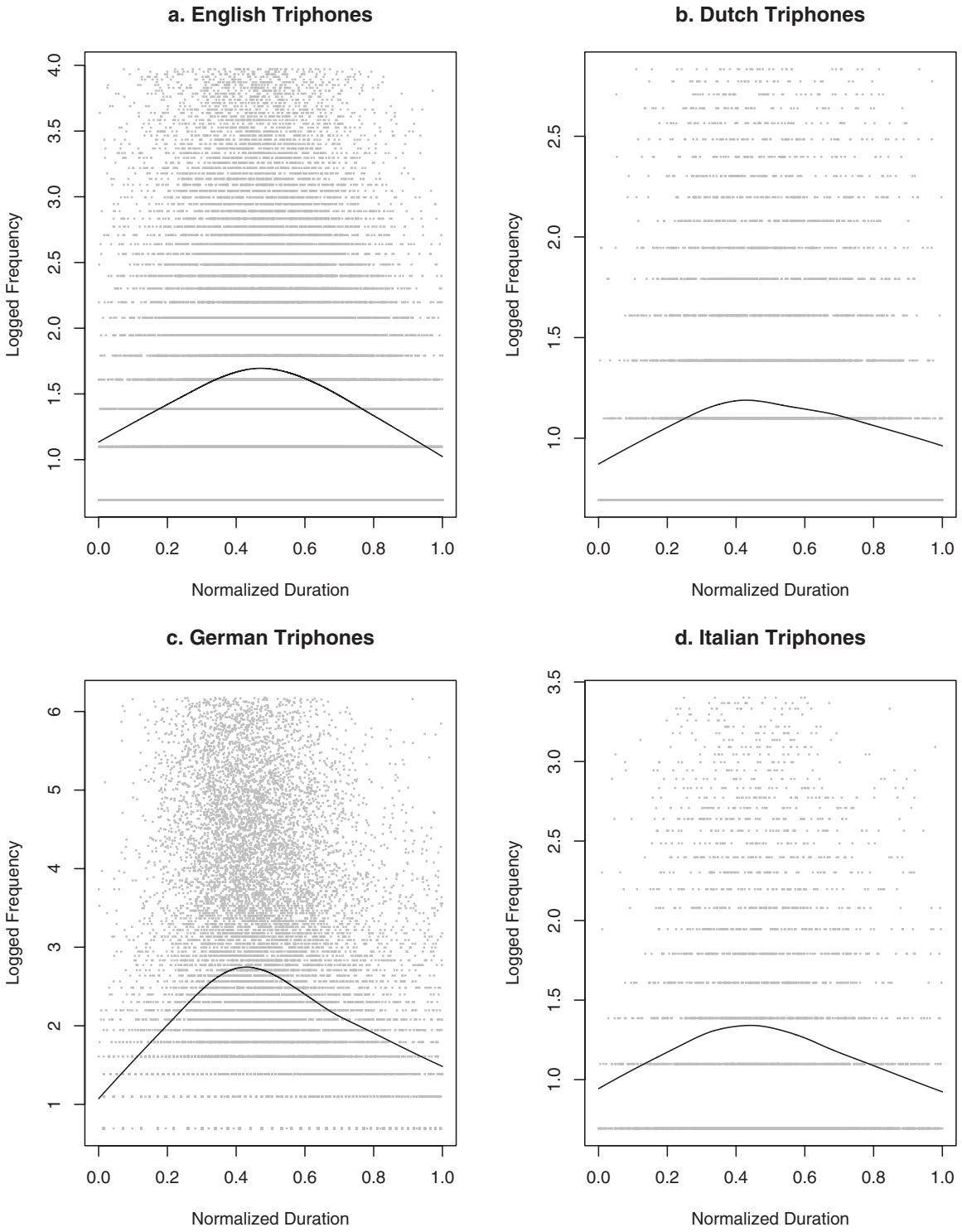

FIG. 2. Distribution of the triphone frequencies over their acoustic durations.

the models. Cook's distance was less than 0.2 for all data points. Table II (triphones) reports the effects of the predictors for frequency.

Duration was a significant predictor of triphone frequency in all datasets as was type, without interactions. Figures 2(a)–2(d) plot the scatterplots for frequency and duration of triphones with addition of the polynomial regression lowess smoother lines.

The nonlinear relations between frequency and duration show concave curves for all four datasets with triphones. As with diphones, the inverse-U shape suggests that speakers tend to avoid phonemic sequences that are either very long or short. Again, this runs counter to the prediction one would make on the basis of Zipf's FPD approach (1935) that frequency should decrease with duration.

IV. VALIDATION OF RESULTS AGAINST ZIPF (1935)

The relationship between the frequency of a speech unit and its acoustic duration can be explored from two view-

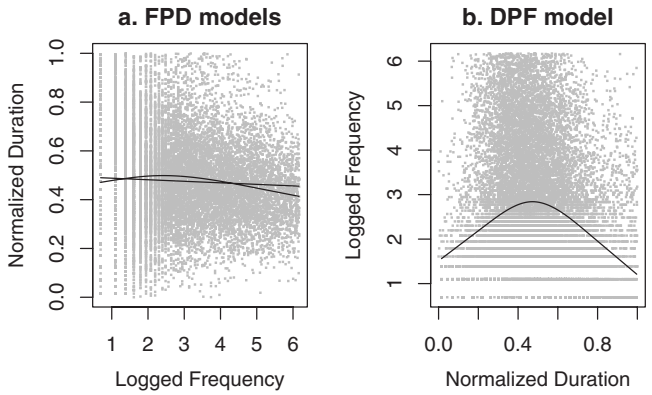

FIG. 3. Function curves of linear and nonlinear FPD models (a) and of the nonlinear DFP model (b) applied to German triphones.

points. In the DPF models that we presented above, acoustic duration predicts frequency of occurrence. In the Zipfian FPD models, the order is reversed: Acoustic duration is the dependent variable, while frequency is considered as an independent variable.

To determine which of the two approaches yields better approximation to the empirical data, we fitted two multiple regression models (DPF and FPD based) to each of the 12 datasets described above. Each model contained only one independent predictor, either frequency or duration, and each correlation with the dependent variable was tested for significant nonlinearities. If the predictors of both the DPF and FPD models reached significance for a given dataset, we identified the best performing model as the model explaining the largest proportion of the variance, R^2. The performances of FPD and DPF models are mathematically identical only if the dependent variable and the predictor show a linear relation.

As an example, Fig. 3 shows the results of the model fitting to the dataset of German triphones. Figure 3(a) displays the scatterplot of duration as a function of frequency (following the Zipfian FPD approach) and plots the linear relation ($R^2=0.008$) as well as the significantly stronger nonlinear relation ($R^2=0.012$) between the two variables. Figure 3(b) swaps the axes in the scatterplot (following our DPF account), plotting frequency as a nonlinear function of duration. The amount of explained variance for this model is 0.04: It thus outperforms both the linear and the nonlinear Zipfian approximations by at least a factor of 3.5.

In the Zipfian models (FPD), frequency of occurrence emerged as a significant linear predictor of acoustic duration for English and Italian uniphones, and as a significant linear or nonlinear predictor for the diphones and triphones of all four languages. Similarly, in the corresponding DPF models, duration reliably predicted frequency.

The DPF and FPD models performed identically for the English and Italian uniphones, which is expected mathematically, given the linear relation between duration and frequency in those two datasets. Crucially, however, for every dataset with diphones or triphones, the amount of variance explained by the FPD model with frequency of use as the independent variable was significantly smaller than the amount explained by the corresponding DPF model, as established by the pairwise comparison of log likelihood ratios of corresponding models. The average R^2 value of the DPF models was 2.6%, while the average R^2 value of the FPD models was 0.2%. DPF models retained their significant advantage over FPD models, when either log durations or log frequencies were z transformed.

Moreover, for the datasets with Italian diphones, English diphones, and German triphones the shape of the function of the Zipfian FPD models is concave. In other words, the shortest elements have the low-to-mid rather than the highest frequency of occurrence. This finding is unexpected in the Zipfian approach. We conclude that Zipf's findings (1935) cannot be extended from uniphones to diphones and triphones and that models with the reverse direction of predictivity (DPF) give rise to qualitatively consistent results (e.g., similar shapes of regression curves) and explain variance in the data better than Zipfian models.

V. CHECKING FOR ARTIFACTS IN THE DIPHONE AND TRIPHONE FREQUENCY DISTRIBUTIONS

Our working assumptions and method of data collection might have given rise to artifacts that produce frequency distribution patterns similar to the ones we observed for the acoustic durations of the diphones and triphones in our datasets. In this section we consider these potential artifacts and demonstrate that none of them can (fully) account for the observed functional relationship of acoustic duration and frequency in the diphones and triphones.

A. Phonotactics

The phonotactics of a language contribute to the frequencies of phonemes. In addition, since phone sequences that violate phonotactic constraints have been shown to facilitate segmentation of continuous speech (e.g., McQueen, 1998), "illegal" n-phones may also be realized longer than legal ones so as to provide better perceptual cues. We set out to validate whether the observed relations between acoustic duration and frequency also hold once the language-specific phonotactic wellformedness of diphones and triphones within words is taken into account. For the diphones and triphones in the Dutch, English, and German datasets we established whether they occur within monomorphemic words [using the CELEX lexical database (Baayen *et al.*, 1995); we did not have access to a corpus of Italian carrying the required information, so this language was left out of consideration]. N-phones that occur within simplex words were coded as "legal," while the others were coded as "illegal." As expected, phonotactically illegal n-phones tended to be longer and less frequent than phonotactically legal ones (for each language, $p<0.0001$). More importantly, the regression analyses replicated the inverse-U concave curves of frequency as a function of acoustic duration in all three languages for both the subset of phonotactically legal and the subset of phonotactically illegal diphones and triphones. The only exception was the English illegal diphones for which a linear function with a negative slope was adequate. We conclude that the inverse-U shaped function predicting frequency from duration is robust with regard to phonotactics.

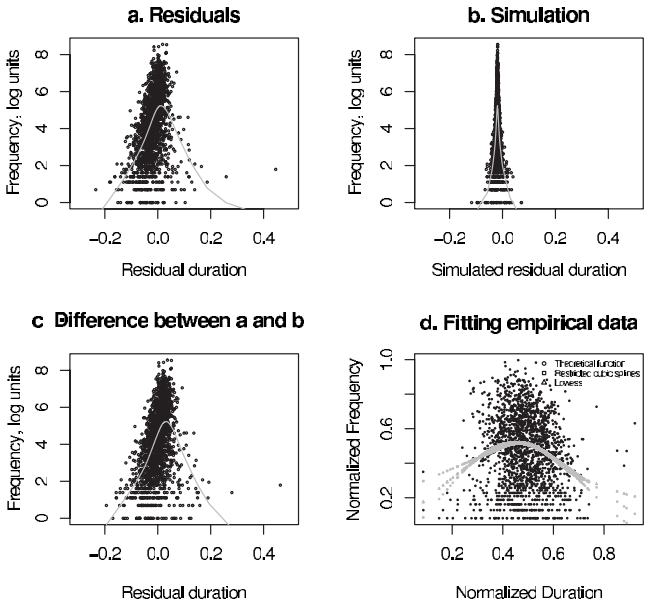

FIG. 4. English diphone frequency as a function of (a) residual mean diphone duration, (b) simulated diphone duration, and (c) the difference between the residual and simulated diphone durations. (d) Approximation of normalized English diphone frequency using the theoretical function (parameter values $a=2.141$, $b=1.737$, and $K=7.447$), the nonlinear regression model using restricted cubic splines, and the added lowess smoother line.

B. Predictors of acoustic duration

Research of the past decades has identified multiple factors that codetermine acoustic duration of n-phones. There is a logical possibility then that the nonlinear relation between acoustic duration and frequency of n-phones is, in fact, a relation between a major predictor of an n-phone's acoustic duration (for instance, word frequency) and n-phone frequency. To test this possibility, we fitted six multiple regression models to the acoustic durations (in milliseconds) of the diphones and triphones in Dutch, English, and German. We only considered n-phones that did not cross word or utterance boundaries. All models included speaker as a random effect to account for intersubject variability in speech rate as well as the following fixed effects: log-transformed word frequency, sum of mean durations of uniphones that constituted the di- or triphone, position of an n-phone in the word and the utterance (both with the levels "initial," "internal," and "final"); and mutual information of the uniphones in the n-phone. The patterns of results were very similar across languages and confirmed the known correlations of these predictors with acoustic duration: n-phones are longer in lower-frequency words, in the beginning and the end positions of both the word and the utterance, if the uniphones they contained were longer, and if the mutual information of these uniphones is larger (all $p's<0.001$). We took the residuals of these models as estimates of acoustic duration from which the effects of these major predictors are regressed out. For all six datasets, we plotted diphone or triphone frequency against the means of those residuals for each n-phone. All resulting plots showed the inverse-U shaped functional relation between the two variables [see Fig. 4(a) for English diphones]. We conclude that the patterns described in Sec. III are unlikely to be artifacts of a dependency between n-phone frequency and one or several factors codetermining acoustic duration of n-phones.

We then compared again the performance of nonlinear FPD (Zipfian) and DPF models now using the mean residualized duration of n-phones instead of the mere mean n-phone duration. Across all subsets of Dutch, English, and German diphones and triphones, our DPF models performed significantly better than the Zipfian models ($p<0.0001$), as indicated by the pairwise comparison of their log likelihood ratios. The average amount of explained variance by the DPF models was 20% as opposed to 11% by the Zipfian FPD models. Thus, acoustic duration is a better predictor of frequency than frequency is of acoustic duration also when the influence of several predictors on acoustic duration is regressed out.

C. Sampling method

A frequency distribution in which extreme values of acoustic duration have the lowest frequency is suspect to the statistical phenomenon of sampling error. An n-phone mean duration will be closer to the grand average duration computed over all n-phones, the more frequent that n-phone is (or, equivalently, the larger the sample size for that n-phone is), since it contributes more to the grand average. That is, less frequent n-phones are predicted to occupy the extreme positions in the distribution of n-phone frequencies over n-phone durations, and more frequent n-phones are predicted to be in the center of that distribution, by virtue of the chosen sampling method. If the number of data points in a population is large enough (like in our datasets), the resulting distribution closely approximates the Gaussian distribution. We investigated whether our sampling method can fully account for the empirical patterns.

We considered the subsets of Dutch, English, and German diphones and triphones, for which mean residual durations were computed (see above). For each of the six subsets we computed the grand average residual duration of all n-phones (μ) and their corresponding SD (σ). If all n-phones were approximately of the same duration, all tokens should together form a Gaussian frequency distribution with the mean (μ) and the SD (σ). For each n-phone in a subset we then took a sample from the corresponding normal distribution (of durations) with μ and σ as parameters, and with the sample size equal to the n-phone frequency. We computed the mean duration for each n-phone sample and plotted it on the x-axis and n-phone frequency on the y-axis to build the simulated frequency distribution of durations. If the observed frequency distributions of the n-phones are just due to sampling error, then the simulated distributions would closely approximate the empirical patterns. We ran 1000 simulations for each subset of diphones and triphones, and we used the Kolmogorov–Smirnov test to estimate the goodness of fit between the simulated and empirical distributions. For all subsets and for all simulations, the Kolmogorov–Smirnov test indicated that the simulated distributions were significantly (all $p's<0.00001$) different from the observed ones. Visual inspection of the simulated distributions [see Fig. 4(b)

for a simulation of the distribution for English diphones] shows that they have a much smaller variance than the empirical ones [shown in Fig. 4(a)]. Furthermore, diphone and triphone frequencies plotted against the *differences* between the observed and simulated durations show the familiar inverse-U shape [see Fig. 4(c)].

We conducted similar simulations using normal distributions with the mean and the SD observed for the specific n-phones (rather than μ and σ of the general data population). Again, for all six subsets and all simulations, the Kolmogorov–Smirnov test showed significant differences between the empirical distributions and the ones simulating random sampling variation. We conclude that the attested inverse-U shapes of n-phone frequency distributions are not artifacts of our sampling procedure.

VI. SELF-ORGANIZATION IN SPEECH

The observed relation between the acoustic duration of an n-phone and its frequency of occurrence may be accounted for by the interacting processes of effort minimization on the part of the speaker as well as on the part of the listener (in some theories of speech production speakers monitor their internal speech via proprioceptive feedback and hence also function as listeners in preferring thorough articulation, cf. e.g., Levelt, 1989). According to the H&H theory (Lindblom, 1990; cf. also Lindblom, 1983; Lindblom, *et al.*, 1984), speakers adaptively balance between the costs of careful speech production and the costs of deficient communication that may come with sloppy pronunciation. This theory has given rise to research on self-organizational properties in speech (De Boer, 2000; Köhler, 1987; Lindblom *et al.*, 1984; Oudeyer, 2005). In what follows, we introduce a theoretical function that quantifies the joint effect of the two minimization processes on n-phone frequencies and we explore how well this function can approximate the observed relation between acoustic duration of an n-phone and its frequency of occurrence.

Several studies have shown that acoustic duration is a measure of ease of speech perception. Longer realizations of speech units tend to facilitate speech comprehension and diminish perceptual confusion (e.g., Janse *et al.*, 2003; Janse, 2004; Kemps *et al.*, 2005; Salverda *et al.*, 2003; but see Ernestus and Baayen, 2007). Acoustic duration is also correlated to ease of speech production. Shorter realizations in general reflect smaller and shorter gestures, which implies less muscular production effort (e.g., Browman and Goldstein, 1992). In line with this notion, Smith *et al.* (1986) demonstrated that subjects show faster production of those uni- and bisyllabic stimuli that a priori were subjectively considered as relatively easy. Likewise, Perkell *et al.* (2002) showed that realizations requiring less articulatory effort (measured as the peak movement speed) tend to be shorter. There are, however, several counterexamples where shorter durations do not always imply easier production (cf., e.g., Beckman and Edwards, 1992; Byrd and Saltzman, 2003). For instance, a shorter duration of a CVC syllable may indicate reduced effort only if it is achieved by shortening its steady state (cf. Nelson, 1983).

Even though the relation between acoustic duration and effort is not straightforward and both articulatory and perceptual complexity are simultaneously affected by many more factors than just duration, we will make for now the simplifying assumption that shorter durations imply minimization of the speaker's articulatory effort and longer durations imply minimization of the listener's perception effort. This assumption will allow us to test how well one can explain the patterns in the empirical data by considering only one inherently noisy dimension of complexity. More specifically, we will investigate to what extent the two opposing tendencies of effort minimization can account for the inverse-U shapes observed in the frequency distributions of n-phones.

We model the tension between these two processes of minimization by considering speech as a dynamic self-regulating system in which a change in the articulatory effort invested by a speaker modulates the effort required of the listener. Both these changes in turn may lead to a change in the frequencies with which speech sounds are used. In what follows, we adopt the framework of Job and Altmann (1985) and Köhler (1987), who modeled the dynamics of sound change as a function of the demands of speech production and comprehension.[1]

The model can be specified in more than one way. For instance, we can model the absolute value of a language property (in this case, n-phone frequency itself, f), or the *amount of change* in n-phone frequency relative to the absolute value of that frequency, df/f. We modeled the relative amounts of change in frequency as we believe that they are more directly influenced by the two opposing tendencies of effort minimization than the frequencies themselves, which are also affected by, for instance, inventory size, phonological generalizations, etc. Furthermore, we assume the simplest relation of direct proportionality between the relative amount of change in frequency and the relative amounts of change in the efforts for the speaker and the listener, $df/f \propto dx$, where x is the total amount of effort defined for both interlocutors. More formally, we hypothesize a complex function of effort $g(x)$ that maps the amount of change in the joint efforts of interlocutors onto the relative amount of change in frequency

$$\frac{df}{f} = g(x)dx. \tag{1}$$

The goal of this modeling exercise is then to specify the functional form of $g(x)$ and validate its goodness of fit against empirical data. Again, we opt for the simplest definitions of our parameters and of the mapping function to test how far these basic assumptions can take us in accounting for patterns observed across four languages.

The speaker's production effort x_s is easier to operationalize than the perception effort of the listener x_l. Here we approximate x_s by the acoustic duration of n-phones. While there is evidence that perception effort strongly correlates with perceptual confusion (e.g., Lindblom, 1990), we remain agnostic as to whether this characteristic is the exhaustive source of effort. To define x_l, we follow Job and Altmann (1985) in making the simplifying assumption that the amount of perception effort is inversely correlated with the amount of production effort, $x_l = 1 - x_s$. This assumption implements

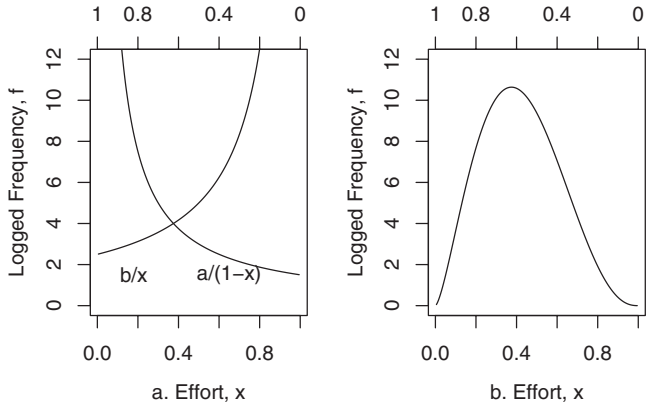

FIG. 5. General shapes of the relation of frequency with articulatory effort (x) and perception effort ($1-x$). (a) Frequency as a function of two processes of effort minimization, separately. (b) Solution for the differential equation (3) with $a=2.5$, $b=1.5$, and $K=150$.

the insight that more careful and thus more effortful articulation alleviates comprehension, while sloppy pronunciation hinders it. We define the variable x_s as the difference between a given amount of effort (an n-phone duration) and the minimal amount of effort (the duration of the shortest n-phone in the dataset), divided by the maximum amount of effort (the duration of the longest n-phone in the dataset). Thus, the value of x_s and of its complement $1-x_s$ are constrained to the interval (0, 1). Since the effort of both the speaker and the listener is now defined in terms of x_s, we henceforth use x to denote x_s and we note that one unit of change is identical for both interlocutors $dx_s=dx_l=dx$.

Recall our hypothesis that the amount of change in frequency relative to the absolute value of frequency is a function of the relative amounts of change in effort for both the speaker and the listener. The amount of change in articulatory effort relative to the absolute value of that effort is given by $g_s(x)dx=b(dx)/x$, where b is a positive coefficient. Likewise, the change in the amount of perception effort for the listener is given by $g_l(x)dx=a(dx)/(1-x)$, where a is a positive coefficient. Figure 5(a) illustrates the situation in which the relative amount of change in frequency in Eq. (1) is only affected by the amount of change in the speaker's effort $g_s(x)dx$, as suggested by Zipf (1935), or only by the change in the listener's effort $g_l(x)dx$. The resulting frequency curves are ideal for either the speaker, or the listener.

Yet we argue that both the speaker and the listener co-determine through their efforts the distributions of n-phone frequencies over n-phone durations. To express the notion of a trade-off between efforts of interlocutors, $g_s(x)dx$ and $g_l(x)dx$, and their joint effect on the change in n-phone frequency, we can model $g(x)dx$ in Eq. (1) either as a difference between the two terms or the division of the two terms. Our further empirical validation showed that the former option provides better fits to observed values; hence, we state that

$$\frac{df}{f} = g(x)dx = (g_s(x) - g_l(x))dx = \left(\frac{b}{x} - \frac{a}{1-x}\right)dx. \qquad (2)$$

When the ideal frequency curves for speaker and listener shown in Fig. 5(a) intersect, the difference between corresponding efforts is equal to zero. In this case, speaker and listener are optimally attuned, and nothing changes in the system. However, if speaker and listener are out of sync, one of the interlocutors has to invest more effort, leading to a difference in the ideal frequencies for the speaker and the listener, and to a change in the likelihood that a given speech sound is used.

The solution of the differential equation (2) is as follows:

$$\log f = b \log x + a \log(1-x) + c, \qquad (3)$$

where c is the constant of integration.

The exponential transformation of Eq. (3) yields the following formula for frequency:

$$f = Kx^b(1-x)^a, \qquad (4)$$

where $0<x_s<1$, and a, b, and K are constants greater than zero.

The curve produced by this function is concave [see Fig. 5(b)] and has its maximum at $x=b/(a+b)$. At this point the frequencies ideal for the speaker and for the listener are equal, and the optimal balance is reached for the system.

The curve is symmetrical if $a=b$. If $a>b$, the maximum shifts leftwards. The area close to the maximum approximates the region of equilibrium where the frequency of a speech sound is least likely to undergo change. In the proximity of the maximum, speakers invest relatively little effort into sound production and at the same time the perceptual efort is relatively low. The position of the equilibrium (and the parameters of this theoretical function) is language specific.

We fitted function (4) to the frequency distributions of uniphones, diphones, and triphones in Dutch, English, German, and Italian using the nls function in the statistical software package R (R Development Core Team, 2007). This program estimated the three constants, a, b, and K, by means of the least squares method. Since the models reported in Sec. III were based on log-transformed values of frequency, we also log-transformed the values of frequency, f, obtained from the theoretical function in Eq. (4). Each dataset was divided into subsets by the levels of CV type, and the theoretical function was fitted to each subset individually. Since our statistical models included CV type as a predictor, splitting of our datasets by CV type was necessary for better accuracy of comparison. Thus, for each uniphone dataset, we obtained two sets of parameters: one that provided the best fit for the vowels and one for the consonants. Similarly, for each diphone dataset, we obtained four such sets, and for each triphone dataset (at most) eight.

The theoretical function did not provide good fits for any of the uniphone datasets. We will therefore only discuss the datasets with diphones and triphones. To estimate the overall goodness of fit, we summed the squared deviations of the fitted values of f from the actual values of frequency over the subsets of each dataset. The resulting sums were then divided by the number of data points in the respective datasets to obtain the mean square errors (MSEs): The smaller the MSE, the closer the fit. We then compared these MSE values with the MSEs of the respective regression models reported above

TABLE III. Estimated parameters of theoretical function. The values in columns, a, b, and K are reported for the models fitted to the logged frequency values of the diphones of the VC type and of the triphones of the CVC type. In column "MSE," the percents in parentheses estimate the performance of the theoretical function as compared to the standard linear regression models for all subsets of given datasets. Thus, −4.3 for the Dutch diphones means that the MSE of the fit to the four subsets of Dutch diphones is 4.3% smaller for the theoretical function than for the regression model.

Dataset	a	b	K	MSE
Dutch diphones	0.27	0.63	0.04	1.47(−4.3%)
English diphones	1.34	0.35	0.00	2.46(0.0%)
German diphones	1.55	0.95	0.00	5.04(−6.1%)
Italian diphones	0.35	0.07	0.08	1.52(0.0%)
Dutch triphones	0.23	0.18	0.19	0.32(−0.1%)
English triphones	0.49	0.17	0.06	0.77(0.0%)
German triphones	0.56	0.46	0.03	2.18(−7.5%)
Italian triphones	0.26	0.14	0.18	0.53(+0.1%)

(estimated as sums of squared residuals divided by the number of data points in the given dataset). The results of the comparison are summarized in Table III.

For the sake of brevity, this table lists the values of the constants for the theoretical function fitted to the diphones of the VC type and to the triphones of the CVC type. The reported MSE values, however, are based on *all* subsets of the datasets. The percents in parentheses estimate the performance of the theoretical function as compared to the standard linear regression models for all subsets of the given datasets. Thus, −4.3 for the Dutch diphones means that the MSE of the fit to the four subsets of Dutch diphones is 4.3% smaller for the theoretical function than for the regression model.

Evidently, the fits to the diphone and triphone data provided by the theoretical function are equivalent to or better than those provided by the standard multiple regression models that use the state-of-the-art approximation of nonlinear functional relations with restricted cubic splines. This is remarkable given that the theoretical function has a predefined shape, which offers less flexibility in fitting than the cubic splines. The two methods are equivalent in the number of parameters they use. The equal or slightly better performance of the theoretical function over regression models using restricted cubic splines also holds when log frequency is normalized and rescaled to the interval between 0 and 1. Figure 4(d) shows fits of the normalized frequency of English diphones of both CV types using the theoretical function and the multiple regression models with restricted cubic splines. We also added as a baseline the fit provided by the locally weighted polynomial regression implemented in the lowess smoother line. The theoretical function based on normalized diphone duration provides a slightly better fit (by 0.4%) to this (normalized) frequency distribution than the multiple regression model.

Since the parameters and coefficients in Eq. (4) are defined in linguistically meaningful terms (the effort of production or perception), this equation affords not only a better fit but also suggests a better interpretability of our findings than the multiple regression models. We conclude that the patterns observed in the frequency distributions of diphones and triphones can be well described by a model that implements the self-regulatory balance in the articulatory and auditory demands of production and comprehension. This strongly suggests that the frequency distributions are codetermined by these two opposing tendencies.

VII. CONCLUDING REMARKS

Across languages, we find significant dependencies between the frequency of occurrence of an n-phone and its acoustic duration. In spontaneous speech in Dutch, English, German, and Italian, speakers prefer diphones and triphones that occupy the middle area of the durational range, and avoid very short durations as well as very long durations. These patterns were consistent across phonetically and phonologically different Germanic languages and a Romance language, which strongly suggests generalizability of our findings and hints that the patterns may derive from fundamental principles of human communication (see Lindblom, 1990). Significant negative correlations were also found between frequency of occurrence and duration of uniphones in English and Italian.

Our approach differs from the approach inspired by Zipf (1929, 1935) in that we predicted frequency from acoustic duration, rather than acoustic duration from frequency. Importantly, multiple regression models based on our DPF approach perform significantly better than the ones that follow the Zipfian approach. This advantage in performance also holds when the influence of several predictors is regressed out of our key factor, acoustic duration. Moreover, the Zipfian account cannot deal with the concave functional form that the relation between frequency and acoustic duration takes under the Zipfian FPD approach.

Essentially, Zipf's approach (1935) is only based on the speaker's tendency to reduce articulatory effort and it correctly predicts that very long n-phones are infrequent. Our data suggest that reduction of comprehension effort may also play a role that becomes evident in speakers' avoidance of very short realizations, which are costly for listeners. We implemented the hypothesis about the interacting demands of efficient speech production and effective speech comprehension mathematically in a theoretical function based on Job and Altmann, 1985. The function provides good fits to the

distributions of frequency of diphones and triphones over their acoustic durations supporting our hypothesis.

Our data point at processes of self-organization in language. Specifically, they document the existence of consistent frequency patterns in several languages, which demonstrate the emergence of global cross-linguistic regularities from the individual instances of communication that operate on a microscopic scale (cf. De Boer, 2001). Clearly, the frequencies of n-phones are determined by the frequencies of words. Changes in the frequencies of n-phones therefore have to result from changes in the pronunciation of words or in word choice, which imply adjustments for the broad linguistic community. The question then is *how* the observed patterns of use spread across vast linguistic communities with such surprising uniformity and in the absence of global control.

Recent computational models connect the emergence of speech sounds with psychologically and socially motivated properties of interactive communication (cf., e.g., De Boer, 2000, 2001, Oudeyer, 2005). We predict for these models that their simulated data will be characterized by inverse-U shaped distributions of sound frequencies over sound durations (similar to the ones we have attested here for four natural languages), probably reflecting the roles of ease of articulation and ease of perception in language use.

ACKNOWLEDGMENTS

The authors wish to thank Alice Turk, Kevin Russell, Austin Frank, and an anonymous reviewer for their valuable comments on previous versions of this manuscript.

[1]One of the alternative approaches considers the amount of information (surprisal) per time unit as a codeterminer of the speaker's and the listener's effort (cf. Aylett and Turk, 2004; Levy and Jaeger, 2006). It argues that communication is optimal (efforts for both interlocutors are minimal) when information density is uniform and close to the capacity of the noisy communication channel. The relationship between present findings and predictions of the uniform information density approach is a topic for further investigation.

Aylett, M., and Turk, A. (**2004**). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," Lang Speech **47**, 31–56.

Aylett, M., and Turk, A. (**2006**). "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllabic nuclei," J. Acoust. Soc. Am. **119**, 3048–3058.

Baayen, R. H. (**1994**). "Productivity in language production," Lang. Cognit. Processes **9**, 447–469.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (**1995**). The CELEX Lexical Database (CD-ROM) (Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA).

Bard, E., Anderson, A., Sotillo, C., Aylett, M., Doherty-Sneddon, G., and Newlands, A. (**2000**). "Controlling the intelligibility of referring expressions in dialogue," J. Mem. Lang. **42**, 1–22.

Beckman, M., and Edwards, J. (**1992**). "Intonational categories and the articulatory control of duration," in *Speech Perception, Production, and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Omaha, Tokyo), pp. 359–375.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., and Gildea, D. (**2003**). "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," J. Acoust. Soc. Am. **113**, 1001–1024.

Scuola Normale Superiore de Pisa (**2001**). *AVIP (Archivio di Varietá di Italiano Parlato)*, [Varieties of spoken Italian archive], edited by P. Bertinetto (Ufficio Pubblicazioni della Classe di Lettere della Scuola Normale Superiore di Pisa, Pisa).

Bolinger, D. (**1963**). "Length, vowel, juncture," Linguistics **1**, 5–29.

Browman, C., and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**, 155–180.

Byrd, D., and Saltzman, E. (**2003**). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," J. Phonetics **31**, 149–180.

Cambier-Langeveld, T. (**2000**). *Temporal Marking of Accents and Boundaries*, (Landelijke Onderzoekschool Taalwetenschap, Amsterdam).

Cleveland, W. S. (**1979**). "Robust locally weighted regression and smoothing scatterplots," J. Am. Stat. Assoc. **74**, 829–836.

Cutler, A., and Clifton, C., Jr. (**1999**). "Comprehending spoken language: A blueprint of the listener," in *The Neurocognition of Language*, edited by C. Brown and P. Hagoort (Oxford University Press, Oxford), pp. 123–166.

De Boer, B. (**2000**). "Self-organization in vowel systems," J. Phonetics **28**, 441–465.

De Boer, B. (**2001**). *The Origins of Vowel Systems* (Oxford University Press, Oxford).

Ernestus, M., and Baayen, R. H. (**2007**). "The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration and frequency of occurrence," in Proceedings of the 16th International Congress of Phonetic Sciences, Saarbruecken, Germany, pp. 773–776.

Fougeron, C., and Keating, P. (**1997**). "Articulatory strengthening at the edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740.

Fowler, C., and Housum, J. (**1987**). "Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction," J. Mem. Lang. **26**, 489–504.

Harrell, F. (**2001**). *Regression Modeling Strategies* (Springer-Verlag, Berlin).

Janse, E. (**2004**). "Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech," Speech Commun. **42**, 155–173.

Janse, E., Nooteboom, S., and Quene, H. (**2003**). "Word-level intelligibility of time-compressed speech: Prosodic and segmental factors," Speech Commun. **41**, 287–301.

Job, U., and Altmann, G. (**1985**). "Ein modell für anstrenungsbedingte lautveränderungen (A model for conditional effort sound changes),"Folia Linguistica Historica **VI**, 401–407.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (**2001**). "Probabilistic relations between words: Evidence from reduction in lexical production," in *Frequency and the Emergence of Linguistic Structure*, edited by J. Bybee and P. Hopper (John Benjamins, Amsterdam).pp. 229–254.

Kemps, R., Wurm, L., Ernestus, M., Schreuder, R., and Baayen, R. (**2005**). "Prosodic cues for morphological complexity in Dutch and English," Lang. Cognit. Processes **20**, 43–73.

Köhler, R. (**1987**). "System theoretical linguistics," Theoretical Linguistics **14**, 241–257.

Ladefoged, P. (**1982**). *A Course in Phonetics*, 2nd ed. (Hartcourt, Brace, Jovanovich, New York).

Levelt, W. J. M. (**1989**). *Speaking. From Intention to Articulation* (MIT, Cambridge, MA).

Levy, R., and Jaeger, F. (**2006**). "Speakers optimize information density through syntactic reduction," in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, (Neural Information Processing Systems Foundation, Vancouver), pp. 29–37.

Lieberman, P. (**1963**). "Some effects of semantic and grammatical context on the production and perception of speech," Lang Speech **6**, 172–187.

Lindblom, B. (**1983**). "Economy of speech gestures," in *The Production of Speech*, edited by P. MacNeilage (Springer-Verlag, New York), pp. 217–245.

Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modeling*, edited by W. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403–440.

Lindblom, B., MacNeilage, P., and Studdert-Kennedy, M. (**1984**). "Self-organizing processes and the explanation of linguistic universals," in *Explanations for Language Universals*, edited by B. Butterworth, B. Comrie, and O. Dahl (Mouton, Berlin), pp. 181–203.

McQueen, J. (**1998**). "Segmentation of continuous speech using phonotactics," J. Mem. Lang. **39**, 21–46.

Nelson, W. L. (**1983**). "Physical principles for economies of skilled movements," Biol. Cybern. **46**, 135–147.

Nooteboom, S. G. (**1972**). *Production and Perception of Vowel Duration: A Study of the Durational Properties of Vowels in Dutch* (University of Utrecht, Utrecht).

Ohala, J. J. (**1996**). "Speech perception is hearing sounds, not tongues,"J. Acoust. Soc. Am. **99**, 1718–1725.
O'Shaughnessy, D., Barbeau, L., Bernardi, D., and Archambault, D. (**1988**). "Diphone speech synthesis," Speech Commun. **7**, 55–65.
Oudeyer, P.-Y. (**2005**). "The self-organization of speech sounds," J. Theor. Biol. **233**, 435–449.
Perkell, J., Zandipour, M., Matthies, M., and Lane, H. (**2002**). "Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues," J. Acoust. Soc. Am. **112**, 1627–1641.
Pitt, M., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (**2005**). "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Commun. **45**, 90–95.
Pluymaekers, M., Ernestus, M., and Baayen, R. (**2005**). "Lexical frequency and acoustic reduction in spoken Dutch," J. Acoust. Soc. Am. **118**, 2561–2569.
R Development Core Team (**2007**). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org (Last viewed 10/1/2008).
Richardson, M., Bilmes, J., and Diorio, C. (**2003**). "Hidden-articulator Markov models for speech recognition," Speech Commun. **41**, 511–529.
Salverda, A., Dahan, D., and McQueen, J. (**2003**). "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," Cognition **90**, 51–89.
Schiel, F., Draxler, C., and Tillmann, H. (**1997**). "The Bavarian archive for speech signals: Resources for the speech community," in Proceedings of the EUROSPEECH 1997, Rhodos, Greece, pp. 1687–1690.
Smith, B., Hillenbrand, J., Wasowitz, J., and Preston, J. (**1986**). "Durational characteristics of vocal and subvocal speech: Implications concerning phonological organization and articulatory difficulty," J. Phonetics **14**, 265–281.
Van Son, R., Binnenpoorte, D., van den Heuvel, H., and Pols, L. (**2001**). "The IFA corpus: A phonemically segmented Dutch open source speech database," in Proceedings of Eurospeech 2001, Aalborg, Denmark.
Van Son, R., and Pols, L. (**2003**). "Information structure and efficiency in speech production," in Proceedings of Eurospeech 2003, Geneva, Switzerland.
Van Son, R., and Van Santen, J. (**2005**). "Duration and spectral balance of intervocalic consonants: A case for efficient communication," Speech Commun. **47**, 100–123.
Zipf, G. K. (**1929**). "Relative frequency as a determinant of phonetic change," Harvard Studies in Classical Philology **15**, 1–95.
Zipf, G. K. (**1935**). *The Psycho-Biology of Language* (Houghton Mifflin, Boston, MA).