Corpus-Based Investigations on the Phonetics of Consonant Voicing*

Bernd Möbius

Abstract

Within and across languages the realization of consonant voicing is highly variable. This study aims to identify, and quantify, the segmental, prosodic and positional factors that have an influence on consonant voicing. A widely used acoustic measure of voicing, viz. voice onset time, is known to have disadvantages both in a cross-linguistic framework, where it fails to provide sufficient information for certain stop consonant classifications, and across consonant classes because it is not defined for fricatives and sonorants. This study applies the voicing profile method to the analysis of voicing properties of consonants in German. The voicing profile is defined as the frame-by-frame voicing status of speech sound realizations in a speech corpus. The speech database was judiciously constructed to cover systematically all possible speech sound combinations in German and a number of positional and prosodic contexts in which these combinations occur. The results are put in a cross-linguistic perspective by comparing the voicing profiles of German stops to those of stops in three other languages, viz. Mandarin Chinese, Hindi, and Mexican Spanish. The results are also discussed in the context of the production and maintenance of voicing during speech production. The voicing profile analysis is intended to serve as a methodology for investigating the discrepancies between the phonemic voicing specification of a speech sound and its phonetic realization in connected speech.

1. Introduction

The realization, or phonetic implementation, of voicing in consonants is highly variable, within and across languages. From the perspective of speech production, voicing can be defined as the presence of (quasi-)periodic vocal fold vibration which produces a (quasi-)periodic excitation signal. Periodicity of the speech signal, a harmonic spectrum, and the presence of low-frequency energy have been identified as acoustic consequences of voicing. Accordingly, the feature [voice] has been defined in articulatory terms (e.g., Chomsky and Halle, 1968; Halle and Stevens, 1971) and in acoustic and auditory terms (e.g., Jakobson et al., 1952; Jakobson and Halle, 1968). More recently, Jessen (2001) has identified 8 acoustic correlates, viz. aspiration duration, closure voicing, fundamental frequency onset, first formant onset, closure duration, preceding vowel duration, following vowel duration, and the difference between the amplitude

Folia Linguistica XXXVIII/1-2 *0165-4004/04/38-5 $ 2.–*

values of the first and second harmonics. He discusses the relevance of each of these parameters for a new, auditorily-based, definition of the features [voice] and [tense], which serve as classification criteria for consonant inventories.

Many studies investigating the acoustic correlates of stop consonant voicing have concentrated on voice onset time (VOT; Lisker and Abramson, 1964), which is defined as the temporal interval between stop release and onset of voicing in the following vowel. It is generally assumed that VOT as a single-dimensional continuous variable can capture the three-way distinction between voiced unaspirated stops (typically with negative VOT values), voiceless unaspirated stops (with VOT values around zero), and voiceless aspirated stops (with large positive VOT values), across languages. The VOT research has been successfully applied cross-linguistically (e.g., Keating, 1984; Ladefoged and Maddieson, 1996; Shimizu, 1990; Poon and Mateer, 1985; Dixit and Brown, 1985). VOT has also been used in automatic speech recognition to improve the identification rate of stops (Niyogi and Ramesh, 2003).

However, the VOT measure is problematic in several respects. For instance, it cannot capture the four-way voiced/voiceless aspirated/unaspirated distinction of stop consonant inventories in languages such as Hindi. Even more importantly, at least for studies of voicing involving not only stops but other classes of consonants too, VOT is not defined in fricatives and sonorants.

To develop a better understanding of the factors affecting consonantal voicing contrasts, the method of constructing voicing profiles was proposed and applied to five languages, viz. German, Mandarin Chinese, Hindi, Mexican Spanish, and Italian (Shih and Möbius, 1998; Shih et al., 1999). This method serves to establish the frame-by-frame probability of voicing throughout the duration of a speech sound realization. It allows to determine the temporal portion of a consonant that is covered by voicing. In the particular case of stop consonants, this is often referred to as "voicing into closure" (Jessen, 2001), whereas a similarly appropriate term appears to be lacking for contextually induced changes of the voicing status of other classes of consonants.

The present paper applies the voicing profile method to the analysis of voicing properties of consonants in German, based on a new speech corpus which was judiciously constructed to cover systematically all possible speech sound combinations in German and a number of positional and prosodic contexts in which these combinations occur. This study aims to identify, and quantify, the segmental, prosodic or positional factors that have an influence on consonant voicing. The voicing profile analysis is intended to serve as a methodology for investigating the discrepancies between the phonemic voicing specification of a speech sound and its phonetic realization in connected speech. The term "phonemic voicing" is used throughout this paper to refer to a canonical (broad phonetic) transcription rather than to an underlying phonological representation;

for instance, the final consonants of the German words *Rad* 'wheel' and *Rat* 'advice' are both represented by [t] in the corpus. Thus, the notational contrast between, e.g., [t] and [d] (as in section 3.1) refers only to cases and positions where the voicing contrast is not canonically neutralized.

The structure of the paper is as follows. Section 2 describes the structure of the speech corpus on which the present investigation is based and the speech processing steps taken to extract voicing information. The process of constructing voicing profiles is also explained in detail. Section 3 presents the voicing profiles obtained for German stop consonants, fricatives and sonorants. These results are put in a cross-linguistic perspective in section 4, where the voicing profiles of German stops are compared to those of stops in three other languages, viz. Mandarin Chinese, Hindi, and Mexican Spanish. The general discussion (section 5) addresses the results of this study in the context of the production and maintenance of voicing during speech production.

2. Database

2.1 Speech corpus

The analysis of consonant voicing reported in this paper has been carried out on a large speech database recorded by one male German speaker. The database (henceforth MS corpus, based on the speaker's initials) was designed for the purpose of unit selection speech synthesis in the SmartKom project (SmartKom, 2003; Schweitzer et al., 2003). SmartKom is a multimodal dialog system designed to perform human–machine dialogs within a number of restricted domains, such as movie theater information, travel planning, and tourist information. The scenario properties entail that speech synthesis in SmartKom must deal with both domain specific and open-domain material.

As a consequence, the speech database was designed to provide not only optimal coverage of domain specific output but also appropriate coverage for the open domain, viz. the entire target language. To this end, sentences were selected from a much larger text database such that they jointly contained a maximum number of distinct combinations of speech sounds and the contexts in which they occur. The core of the speech corpus comprises a full set of diphones, but beyond this core the corpus also contains rich combinations of phones and contextual factors, including segmental context, syllable structure, and syllabic stress, as well as positional and intonational factors. Information on the size and structure of the MS corpus, as far as it is relevant for the present study, is given in Table 1.

Length	Sentences	Words	Consonants	Speakers
160 min.	2601	17489	56434	1

Table 1: Size and structure of the MS speech corpus.

The MS speech corpus was automatically segmented on the phone, syllable and word levels by HMM-based forced alignment (Rapp, 1995). Diagnostic tools were used to detect gross segmentation errors, which were then manually corrected; other segmentation errors were manually corrected whenever they were found, but a systematic manual inspection of automatic segmentation results was impractical due to the size of the speech database.

The text and prosodic analysis components of the IMS German Festival text-to-speech system (IMS Festival, 2003) were used to compute a hierarchical symbolic corpus annotation. The prosodic annotation was manually checked and, if necessary, corrected. For each speech sound exemplar in the corpus a feature vector was constructed that includes (i) its phonemic identity, (ii) its position in the syllable (onset or rhyme), (iii) the phonemic identity of its left and right neighbors, (iv) presence or absence of syllabic stress on the corresponding syllable, (v) type or absence of pitch accent on the syllable, (vi) type or absence of boundary tone on the syllable, (vii) position of the syllable in the phrase (initial, medial and final), and (viii) word class of the related word (function word or content word).

In the present study, these features were exploited as potential factors affecting the voicing properties of speech sounds. All statistical analyses were performed by means of the statistics software package R (R Project, 2001).

2.2 Speech processing

Voicing information was obtained automatically by means of the "get_f0" program included in the ESPS/xwaves speech analysis software, version 5.3 (Entropic Inc.). The program reports a binary voicing decision for each analysis frame (10 ms steps), with "1" indicating "voiced" and "0" indicating "unvoiced". It is possible to introduce a preference factor in the voiced/unvoiced decision to encourage either the voiced or the unvoiced hypothesis, if properties of the acoustic signal, for instance periodic background noise, or characteristics of the speaker's voice are known to bias the algorithm. Given the fact that the MS corpus was recorded in an anechoic chamber, the signal quality was considered to be of no concern; since no indication of a voiced/unvoiced bias was detected, the default (neutral) preference factor was applied. The frame-by-frame binary classification output of "get_f0" was therefore taken as the raw data for further analysis.

2.3 Voicing profiles

For each consonant exemplar in the speech database, 11 samples of voicing information were obtained at 10 equidistant time intervals through the duration of the segment (Figure 1). Because of the granularity of the automatic segmentation (see below), the initial and final analysis frames of each speech sound were excluded, yielding voicing information at 9 time points, viz. at 10%, 20%, ..., 90% of the duration of each segment.

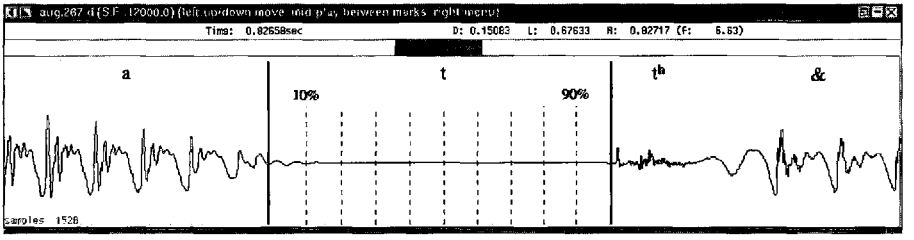

Figure 1: Voicing information is sampled at 10 equidistant time intervals through the duration of each consonant exemplar (here [t] closure) in the corpus.

The voicing probability of a speech sound at a given temporal position is computed as the percentage of the population, i.e. realizations (or exemplars) of the speech sound in the corpus, that is voiced at that position. We refer to this frame-by-frame voicing status of a speech sound as its voicing profile. Voicing profiles show the dynamic changes of voicing probability of a given speech sound as a function of normalized time. In the voicing profiles, voicing probability is plotted on the y-axis as a function of 9 normalized time points, the latter shown as percentage of the consonant duration. For example, Figure 2 shows that 43% of the realizations of the voiceless alveolar stop [t] in German are voiced at the beginning of the closure phase and 4% are still voiced near the end of the closure phase.

When applied to stop consonants, the voicing profile method is meaningful only for the closure phase. It is therefore of great importance to have precise temporal information on the start and end of the closure. The automatic alignment tool used to obtain segmentation information does not split stop consonants into their closure and release phases but instead marks the start and end of the entire stop consonant. Fortunately, the aligner usually places the end-of-segment mark right after the plosive burst (Figure 3). Given the fact that the burst is a very short event that even in the case of voiceless stops hardly ever exceeds 10 ms, it is fair to say that the imprecision introduced by including the burst into the stop consonant duration is of approximately the same order of magnitude as the aligner's frame step width, and therefore does not constitute an additional source of uncertainty for the analysis.

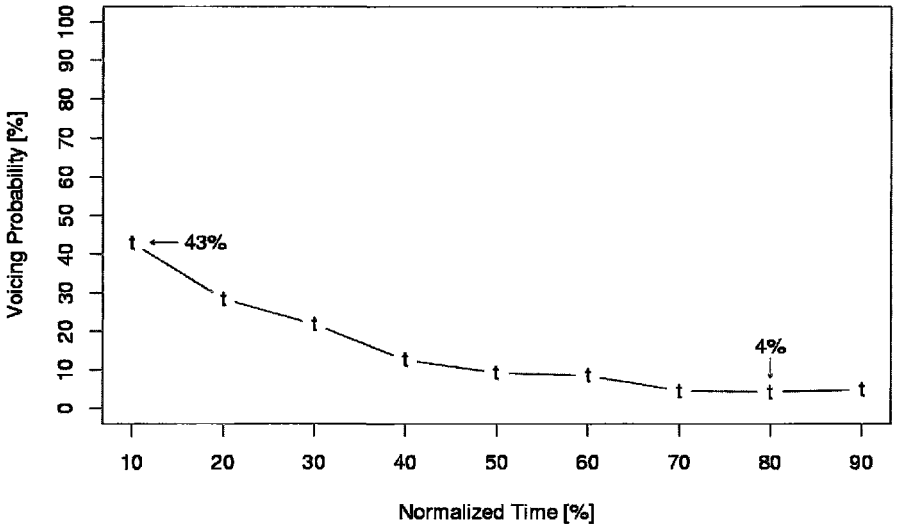

Figure 2: Voicing profile of German [t]. 43% of the [t] realizations in the corpus are voiced at the beginning of the closure phase and 4% are still voiced near the end of the closure.

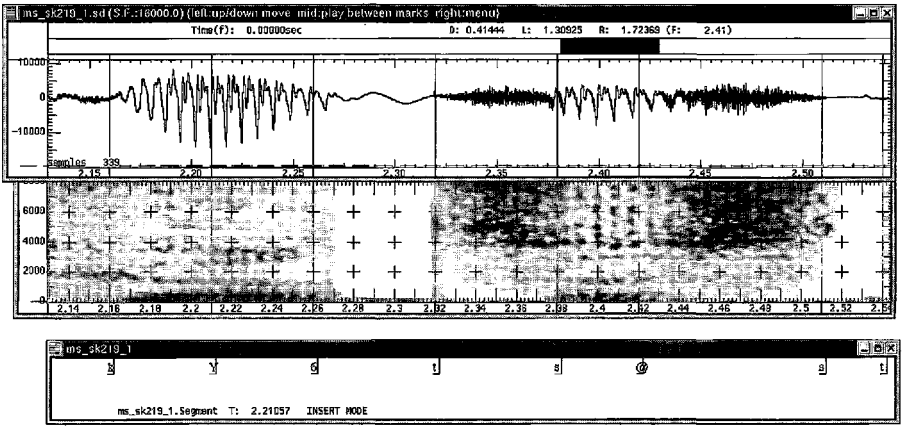

Figure 3: Example of the automatic aligner's performance.

The boundary between the speech sound preceding a stop and the beginning of the closure phase is marked by the aligner at approximately the same place as human labelers usually mark it. For instance, the boundary between a vowel or sonorant consonant and the following stop consonant is consistently placed in the region where the vowel formants, especially the second formant, disappear (Figure 3), which is a good indicator that the stop closure is being formed.

There is a tendency for the aligner to place the boundary between a fricative and a preceding vowel or sonorant consonant slightly earlier than a human

labeler would do. In Figure 3 the ideal boundary location between [ə] (SAMPA symbol [@]) and [s] is arguably one fundamental period to the right of where the aligner placed it. Again, the temporal difference between the actual and the ideal boundary location is no more than one analysis frame.

To alleviate any concern with respect to artifacts resulting from the automatic placement of segmentation marks, the initial and final analysis frames were excluded from the analysis. As the presentation of the voicing profiles in the following section will demonstrate, the observed context effects on consonant voicing extend well beyond the initial frames; they are often measurable throughout most or all of the duration of the segment and cannot be attributed to segmentation inaccuracy.

3. Voicing profiles in German

This section presents the voicing profiles of consonants in German. Table 2 displays the inventory of German consonants along with their frequency of occurrence in the MS corpus. Note that the numbers of individual consonants do not add up to the total number of consonants (56434) given in Table 1 because the corpus also includes realizations of foreign-language speech sounds, which were excluded from the present study due to their relatively low frequency and imbalanced contexts. Results for stop consonants are presented in section 3.1, followed by those for fricatives (section 3.2) and sonorants (section 3.3).

p 1409	t 7996	k 2353	b 1800	d 3611	g 1669	ʔ 4905		
f 2385	s 4593	ʃ 1549	ç 1181	x 518	h 968	v 1539	z 1654	ʒ 162
m 2708	n 8285	ŋ 669	l 3329	R 2472	j 472			

Table 2: German consonant inventory, with frequencies of occurrence in the MS corpus.

3.1 Stop consonants

Figure 4 shows the voicing profiles of the closure phases of German stops, pooled across all left and right segmental contexts. The most robust effect is the neat separation of the two stop series, viz. phonemically voiced and phonemically voiceless stops. The probability of voicing in the [b,d,g] series is consistent across the duration of the closure phase and stays within a narrow range (approximately 60-75%). Note that the voicing probability nowhere approaches 100%, not even at the beginning of the closure phase. The [p,t,k] series, on the

other hand, shows a considerable degree of voicing early in the closure, but the probability of voicing falls to under 10% by the temporal mid point of the closure.

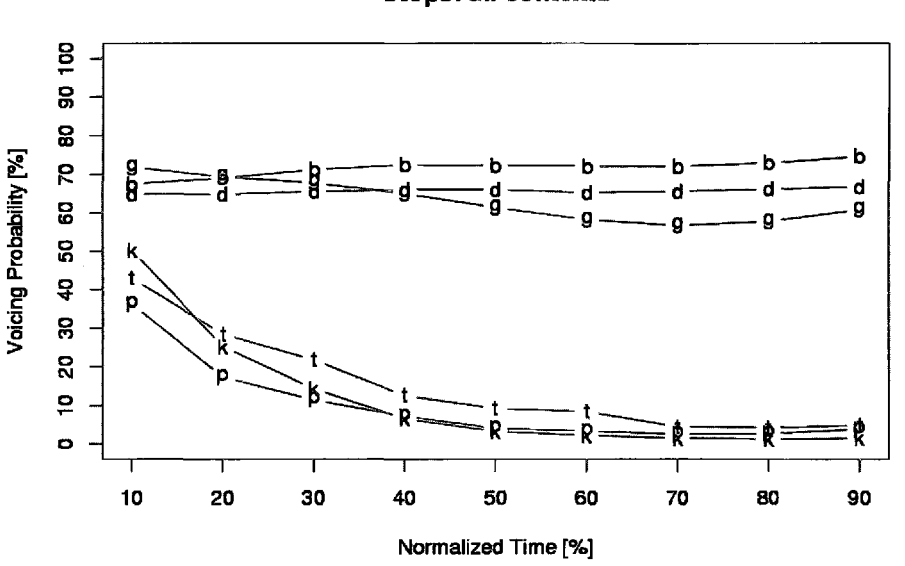

Figure 4: Voicing profiles of the closure phases of German stop consonants, pooled across all left and right segmental contexts.

Two main effects may be inferred from the voicing profiles in Figure 4: an overall devoicing effect on the phonemically voiced stops and a closure-initial voicing effect on the phonemically voiceless stops. To disentangle these effects, separate voicing profiles were constructed for two types of left-hand segmental context. Figure 5 shows the closure voicing profiles for vocalic or sonorant left contexts (solid lines) and for voiceless obstruent left contexts (dashed lines).

In the case of the phonemically voiced stops, the type of left context is evidently the main factor affecting the voicing probability of the entire closure phase, because the entire [b,d,g] voicing profiles are shifted upward from the 60-75% range to approximately 90% for [b] and [d], and around 80% for [g], when the left context is a sonorant consonant or a vowel. For voiceless obstruent left contexts, the voicing probability stays below 10% for almost the entire closure phase. This strong devoicing effect is all the more remarkable as German phonotactics requires a syllable boundary to be present between a phonemically voiced stop and a preceding obstruent. Evidently, the syllable boundary does not act as a strict coarticulation boundary here.

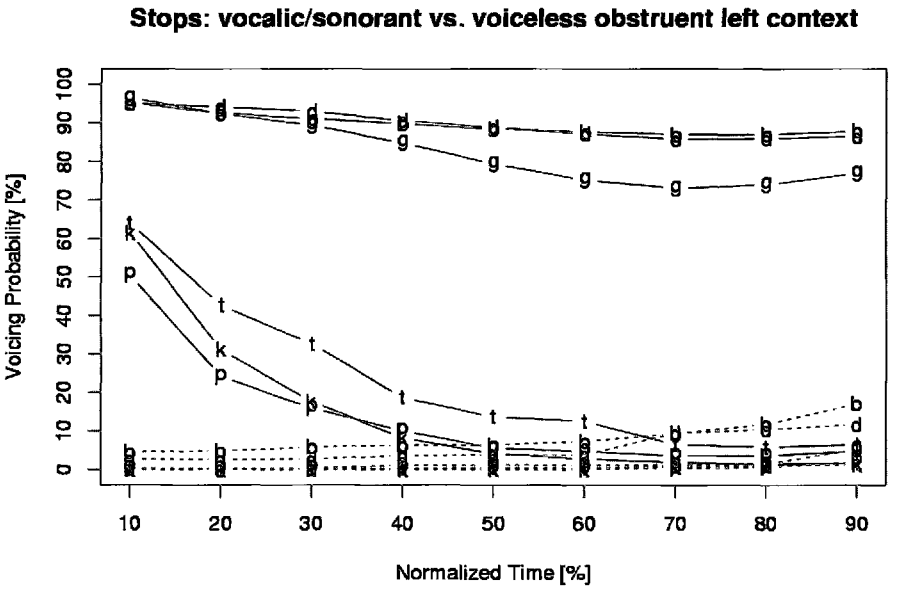

Figure 5: Voicing profiles of closure phases for vocalic or sonorant left contexts (solid lines) and for voiceless obstruent left contexts (dashed lines).

In the case of the phonemically voiceless stops, the overall shape of the voicing profiles remains unchanged across left segmental contexts. As one would expect, a vocalic or sonorant context raises the probability of voicing for [p,t,k] by approximately 10-15% in the first half of the closure phase. For voiceless obstruent left contexts, the probability of voicing is practically zero.

A relatively weak right-context effect can be observed too. The [b,d,g] voicing profiles in voiceless obstruent left contexts (dashed lines) in Figure 5 show a slightly rising probability of voicing towards the end of the closure, which may reflect an occasionally occurring prevoicing, in anticipation of the right segmental context, where according to German phonotactics only sonorant consonants or vowels can occur. The same effect is discernable indirectly in the [b,d,g] voicing profiles in vocalic or sonorant left contexts (solid lines), which should be expected to continue to fall throughout the closure phase in the (hypothetical) absence of the vocalic or sonorant right context.

3.2 Fricatives

Figure 6 displays the voicing profiles of German fricatives, pooled across all left and right segmental contexts. The phonemically voiceless fricatives ([f,s,ʃ,ç,x], SAMPA symbols [f,s,S,C,x]) show a considerable amount of initial voicing. For [f,s,ʃ] the voicing probability drops below 10% after about one third of the

respective speech sound's duration. The probability of voicing is quite high early in [ç] and [x]; whereas for [ç] voicing practically disappears by the temporal mid point, some voicing probability remains throughout the duration of [x].

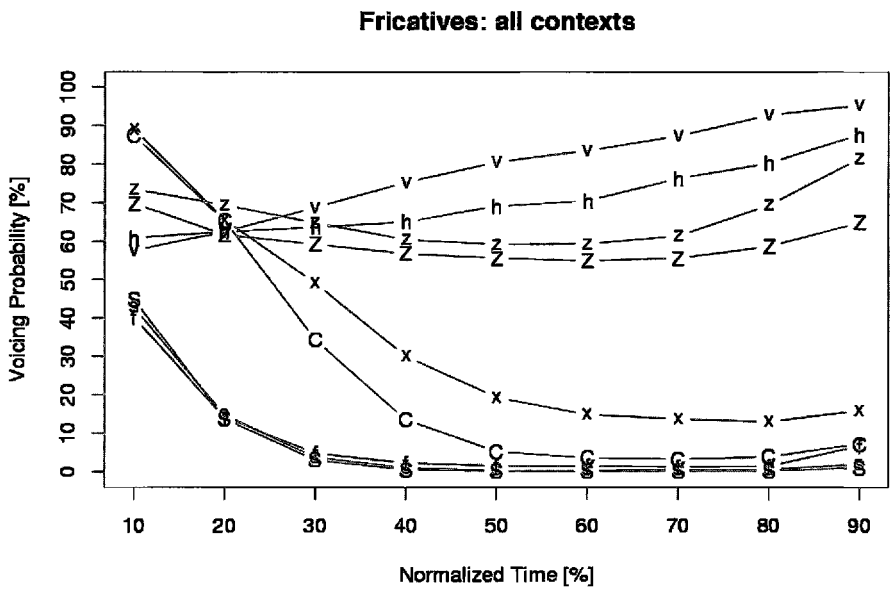

Figure 6: Voicing profiles of German fricatives, pooled across all left and right segmental contexts.

The voicing probability of the phonemically voiced fricatives [z] and [ʒ] (SAMPA symbol [Z]) stays within a rather narrow range (approximately 60-80%), similar to what was observed for the phonemically voiced stops, but with a minimum in the voicing profile around or shortly after the temporal mid point. Again, voicing probability nowhere approaches 100%, not even at the beginning of the speech sound. The voicing probability of [v] rises monotonously throughout the speech sound's duration, from just under 60% to a final value of 95%. Interestingly, the fricative [h] patterns with the phonemically voiced fricatives and most closely with [v]. In most text books on German phonetics and phonology (e.g., Kohler, 1995, and Pompino-Marschall, 1995; cf. the discussion of the feature specification of [h] in Wiese, 2000), [h] is classified as a voiceless fricative, even though the high probability of voiced [h] realizations in all-voiced segmental contexts has been known for some time (e.g., Stock, 1971).

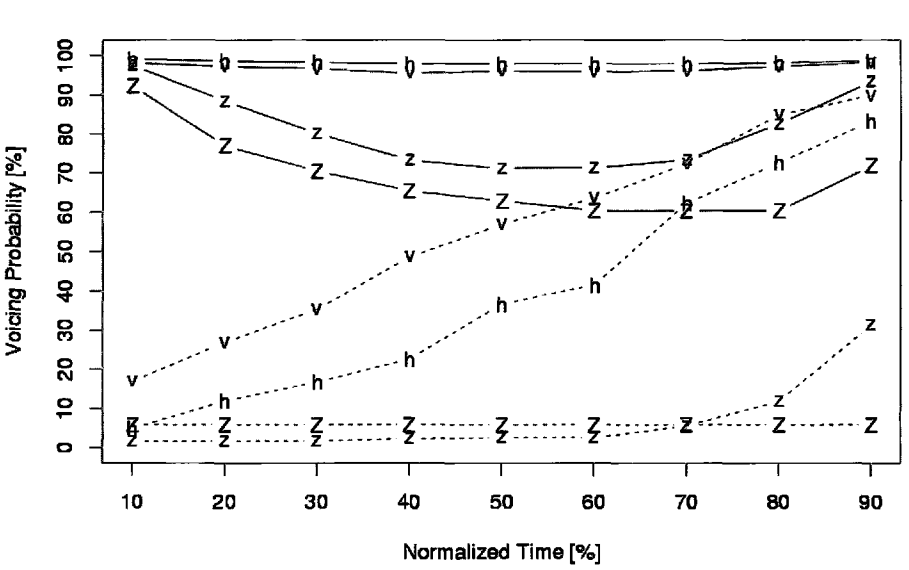

Figure 7: Voicing profiles of phonemically voiced fricatives and [h], separated by type of left segmental context: vocalic or sonorant (solid lines) vs. voiceless obstruent (dashed lines).

An interesting picture emerges when separate voicing profiles are constructed for the phonemically voiced fricatives depending on the type of left-hand segmental context (Figure 7). The parallel patterns of [v] and [h] become even more evident: in the vocalic or sonorant left context, the two fricatives are almost perfectly voiced throughout, whereas in a voiceless obstruent left context they both rise monotonously from an initially very low voicing probability to about 80-90% voicing probability near the end of the speech sound. For [z] and [ʒ] we observe the pattern that was already visible in Figure 6, viz. the minimum in voicing probability shortly after the temporal mid point of the speech sound in vocalic or sonorant left contexts. In a voiceless obstruent left context, both fricatives are practically voiceless throughout (but note that there are too few exemplars of post-obstruent [ʒ] in the corpus for its respective voicing profile to be representative). Due to the phonotactics of German, the right context for the cases displayed in Figure 7 is almost invariably a vowel (there are only two exemplars of a [vl] onset in the corpus and none for [vr]).

Splitting the left-hand contexts of the voiceless fricatives does not add much information to what is already evident in Figure 6. The only noteworthy difference is an increased probability of voicing (about 70%, as opposed to 45% when contexts are pooled) in the initial frames of [s] and [ʃ], and 55% as opposed to 40% for [f], followed by a sharp drop to a negligible degree of voicing after the

first 30% of the respective speech sound's duration. The voicing profiles of [ç] and [x] do not change significantly.

To further explore the conspicuous drop in voicing probability shortly after the temporal mid point of [z] and [ʒ] even in all-sonorant contexts, we considered the absolute duration of the respective speech sound as a possible factor. Initiating and maintaining vocal fold vibration is a delicate process during which several articulatory gestures need to be coordinated, among them appropriate sublaryngeal and supralaryngeal pressure ratios and vocal tract enlargement (Jessen, 2001). The constriction of the vocal tract required to produce a (voiced) fricative causes adverse aerodynamic conditions for vocal fold vibration, and the longer the constriction is maintained the more likely the voicing process is to break down.

Figure 8 illustrates the effect of fricative duration on the probability of voicing for [v] (dashed line) and [z] (solid line). The voicing profiles are coded according to the absolute duration of the fricative realizations in the corpus: (A) <60 ms, N=306; (B) 60-80 ms, N=670; (C) 80-100 ms, N=484; (D) >100 ms, N=194. After about one third of the [z] duration there is a clear negative correlation between the absolute duration of the speech sound and its voicing probability. The effect becomes stronger as the absolute duration increases; for [z] realizations with a duration of more than 100 ms the probability of voicing is less than 60% after the temporal mid point.

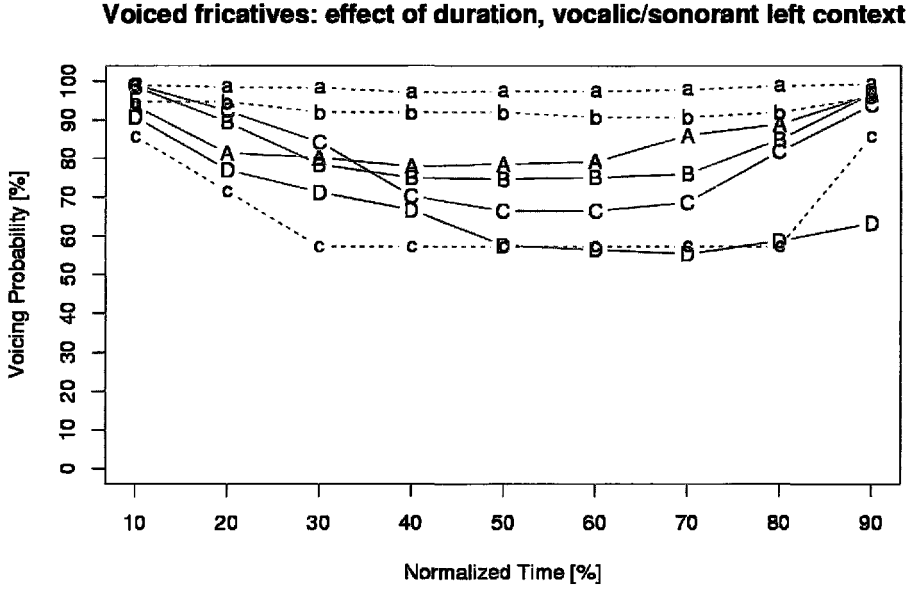

Figure 8: Effect of absolute duration on the voicing probability of [v] (dashed lines) and [z] (solid lines).

For [v] the effect is similar but not identical. Realizations of [v] with a duration of less than 80 ms tend to stay voiced throughout, but [v] exemplars with a duration of more than 80 ms have the same lowered probability of voicing as the longer [z] exemplars do. The voicing profile for [v] realizations longer than 100 ms is omitted because of sparse data (N=6).

3.3 Sonorants

Similar to the pattern observed for the phonemically voiced fricatives, the type of left-hand segmental context is the single most important factor affecting the voicing properties of sonorant consonants in German. Figure 9 displays the voicing profiles of the sonorants [m,n,ŋ,l,R,j]; for the purpose of this study the glide [j] is included in the class of sonorants, but notice that its status as either a sonorant or a fricative or a non-syllabic vowel in German is controversial (see discussion in Wiese, 2000). In the vocalic or sonorant left context (solid lines), all sonorants are practically fully voiced throughout their duration, with the minor exception of [R] which has some voiceless exemplars in the corpus.

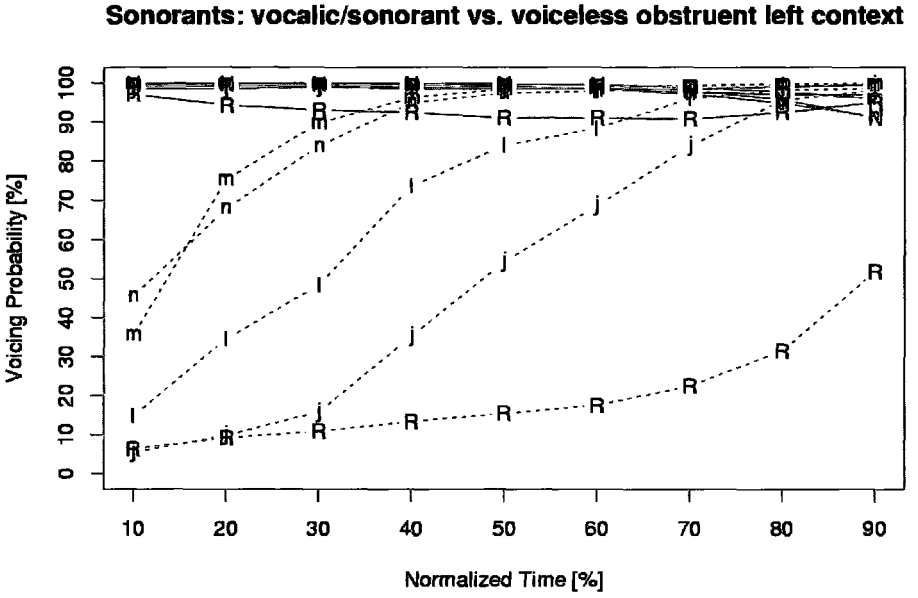

Figure 9: Voicing profiles of sonorants, separated by the type of left segmental context: vocalic or sonorant (solid lines) vs. voiceless obstruent (dashed lines).

The picture changes in interesting ways when the sonorants are preceded by a voiceless obstruent. In general, all sonorants undergo initial devoicing, but the extent differs between speech sounds. For instance, the nasals [m] and [n] ([ŋ] is phonotactically impossible in this position) rise from a probability of voicing

of around 40% to almost perfect voicing after one third of their respective duration, whereas [l] and [j] tend to rise much more slowly from initially mostly voiceless frames to high degrees of voicing only near the end of the respective speech sound's duration. Note that the data are pooled for all right-hand contexts, although due to phonotactical constraints voiced contexts are dominant (sonorants in the syllable onset can only be followed by a vowel, and [j] and [R] can only occur prevocalically anyway).

A special case is [R], which tends to be voiceless throughout when preceded by a voiceless obstruent. Speaker MS usually produces a velar or uvular voiced fricative [R] realization, which in its devoiced variant is virtually indistinguishable from [x] when played in isolation. From this point of view it would be appropriate to subsume [R] under the set of phonemically voiced fricatives above, especially since its voicing profiles under the two left-context conditions are similar to those of [z]. A parallel alternation between sonorant and fricative is observed for [j], which tends to attain [ç]-like frication when preceded by a voiceless obstruent.

The strength of the devoicing effect may depend on the presence or absence of a syllable boundary between the sonorant and the voiceless obstruent preceding it. To investigate this question, separate voicing profiles were constructed for obstruent-sonorant sequences with and without an intervening syllable boundary.

Figure 10 displays the results for the voiceless obstruent left context with syllable boundary (solid lines), as in *Stecknadel* [ʃtɛk.na:dəl] 'pin', and for the same segmental context without syllable boundary (dashed lines), as in *knapp* [knap] 'tight'. The effect is consistent across all sonorants: the probability of voicing starts from, and remains on, a higher level if a syllable boundary separates the obstruent-sonorant cluster; otherwise the paired profiles for each sonorant are almost perfectly parallel. The weakest effect is observed for the nasals, and the strongest for [j] and [R], as one would expect based on their realizational alternation with fricatives. Taken together, a syllable boundary tends to weaken the effect of sonorant devoicing exerted by a preceding voiceless obstruent, but it does not act as a strict coarticulation boundary.

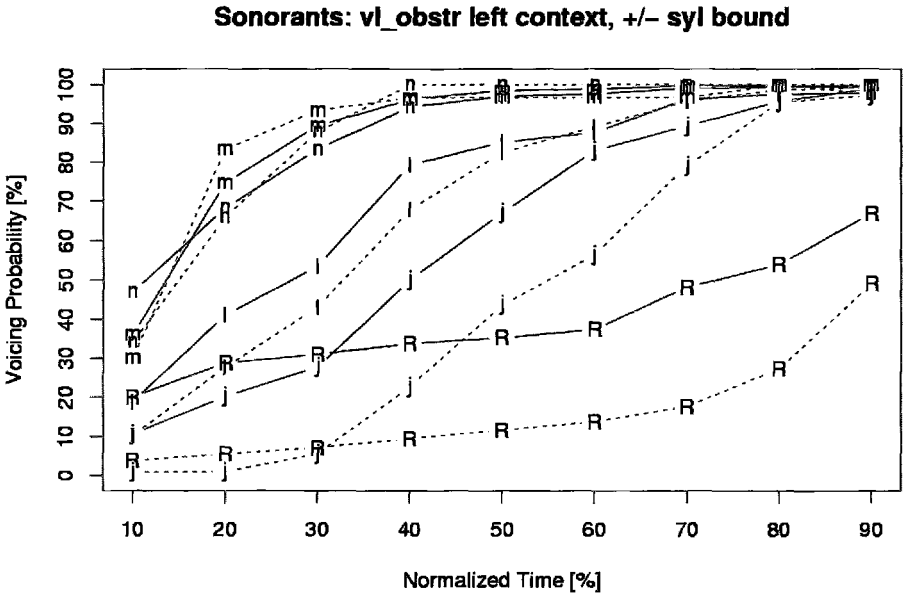

Figure 10: Effect of a syllable boundary between a sonorant and a preceding voiceless obstruent: presence (solid lines) vs. absence (dashed lines) of syllable boundary.

4. Cross-linguistic comparisons

The voicing profile methodology has been applied successfully to languages other than German. To put the results for the voicing profiles of German consonants in a cross-linguistic perspective, we compare the voicing profiles of German stops to those of stops in three other languages, viz. Mandarin Chinese, Hindi, and Mexican Spanish. Data for these languages are taken from our previous studies (Shih and Möbius, 1998; Shih et al., 1999), where further details on voicing profiles in these languages can also be found.

The most robust effect on German stop closure voicing reported in section 3.1 (Figure 4) is the neat separation of the two stop series, viz. phonemically voiced and phonemically voiceless stops. In Mandarin, on the other hand, all stops are phonemically specified as voiceless; the main contrast is defined in terms of aspiration. In Figure 11 [p,t,k] are used as symbols representing voiceless unaspirated stops, and [P,T,K] for voiceless aspirated stops.

The most robust effect found for Mandarin stops is the separation of unaspirated and aspirated sounds. Unaspirated stops are more likely to be voiced than aspirated ones. This result is consistent with the observation that in aspirated stops, vocal fold abduction is required to build up pressure in preparation of the aspiration phase after stop release, thereby effectively preventing vocal fold vibration during the closure phase (Jessen, 1999). The segmental context has

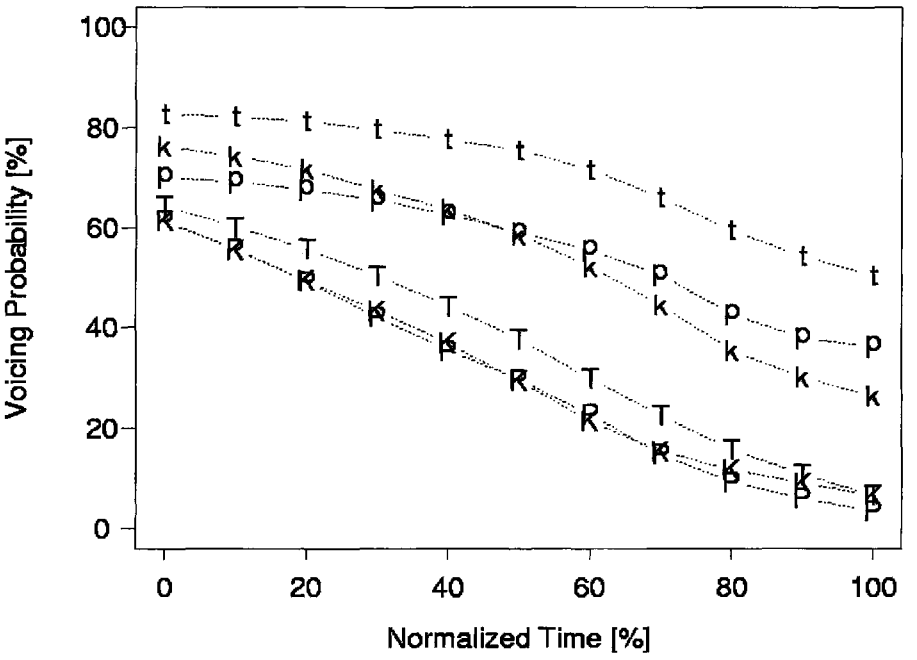

Figure 11: Voicing profiles of Mandarin Chinese stop consonant closures. Adapted from Shih and Möbius, 1998.

little effect on stop closure voicing due to distributional restrictions in Mandarin: stops can only occur in syllable onsets, there are no consonant clusters and no coda obstruents; taken together, stops are always surrounded by voiced sounds.

The results show that, despite the typological difference between the consonantal systems of the two languages and despite differences in the complexity of syllable structure, Mandarin [p,t,k], phonemically voiceless and unaspirated, and German [b,d,g], phonemically voiced and unaspirated, show very similar patterns in their voicing profiles. This result supports an analysis which favors the feature [spread glottis] (or [aspirated] or [tense]) over [voice] as the primary feature for distinguishing the two series of stops, not only in Mandarin but in German as well (e.g., Jessen and Ringen, 2002).

Hindi has a large inventory of stops contrasting in voicing, aspiration and four places of articulation. The voicing profiles displayed in Figure 12 show that the stop populations are divided by voicing. Phonemically voiced stops tend to be fully voiced, especially in the center region, while phonemically voiceless stops turn voiceless after about one third of their duration. Thus, the phonetic voicing property of Hindi stop closures corresponds well with the phonemic specification of the stops.

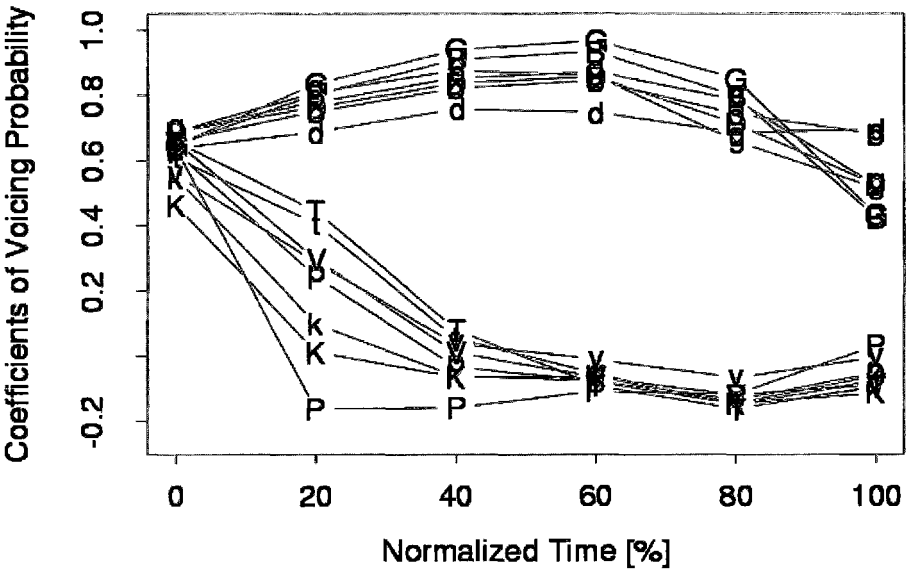

Figure 12: Voicing profiles of Hindi stop consonant closures. Adapted from Shih et al., 1999.

Aspiration has a significant effect on voicing. For instance, voiced aspirated stops (upper-case symbols in Figure 12) are more solidly voiced than their unaspirated counterparts (lower-case symbols) in the center region. Moreover, aspiration has a weak devoicing effect on phonemically voiceless stops: in aspirated/unaspirated pairs, the aspirated stop tends to have a lower probability of voicing at a given position or, to put it differently, an aspirated voiceless stop becomes voiceless earlier than the corresponding unaspirated one, everything else being equal (Shih et al., 1999).

An almost perfect correspondence between phonemic specification and phonetic properties is found in Mexican Spanish (Shih et al., 1999), where the voiced stops ([b,d,g]) are clearly separated from the voiceless stops ([p,t,k]) and the voiceless alveolar affricate [C] (Figure 13). The differentiation is perfect near the end of the closure phase but much less so at the beginning, where the effect of the segmental context is significant. As observed with the other languages, the voiceless stops are likely to have voicing that extends far into the closure. A very similar pattern was found for stop closure voicing in Italian (Shih et al., 1999).

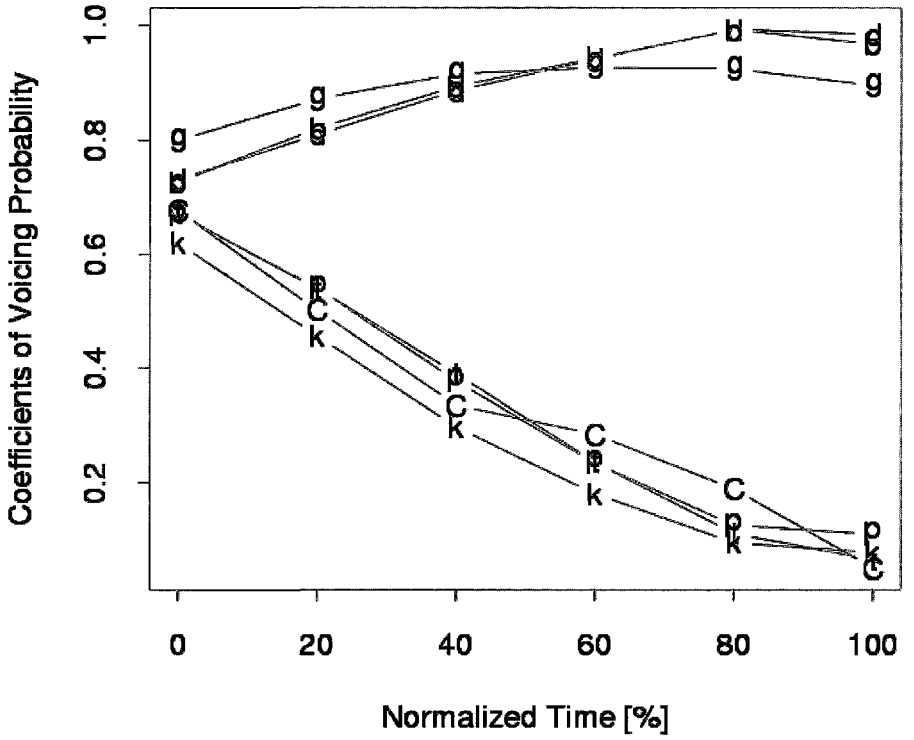

Figure 13: Voicing profiles of Mexican Spanish stop consonant closures. Adapted from Shih et al., 1999.

5. General discussion

The voicing profile analysis serves as a methodology for investigating the discrepancies between the phonemic voicing status of a speech sound and its phonetic realization in connected speech. For instance, stop consonants in the languages under investigation here and in a previous study (Shih et al., 1999) can be classified by different combinations of the features [voice] and [aspirated]. Our results show that, despite the differences in phonemic specification, Mandarin voiceless unaspirated stop closures show voicing profiles similar to the voiced unaspirated stop closures in German. Similarly, the voiceless aspirated stops of Mandarin pattern with the voiceless (aspirated or unaspirated) stops in German, Mexican Spanish, and Italian. Hindi has a two-way phonemic distinction between voiced/voiceless and aspirated/unaspirated stops. Hindi stops can be ranked along a scale of decreasing probability of voicing, from voiced-aspirated to voiced-unaspirated to voiceless-unaspirated to voiceless-aspirated, e.g.: [G] > [g] >> [k] > [K] (Shih et al., 1999).

Our research suggests that a binary specification of voicing over the domain of the entire speech sound is often insufficient to differentiate the stop series in a given language. It may also obscure similarities or parallel patterns across languages. The VOT measure has been argued to provide a better classification of stops. However, in Mandarin and German the two populations of stops are differentiated by the patterns of voicing in the closure phase. Voicing profiles, as suggested in our study, allow us to describe the dynamic changes of the voicing status of speech sounds as a function of (normalized) time. In the conventional usage of VOT, voicedness is expressed as negative VOT counting backward from the time of the stop release. Since voicing typically ceases before the burst in all stops of Mandarin, German and Hindi, the more voiced and the less voiced populations in these languages cannot be differentiated by negative VOT alone.

The important role of contextually induced voicing, especially for the characterization of stops, requires a high precision of speech sound segmentation. This is of concern particularly in the case of corpora that have been segmented automatically. We therefore compared the voicing profiles of German stop consonants obtained from two different corpora and thus two different speakers.

The voicing profiles of stops in the MS corpus are indeed quite similar to the ones obtained from a speech corpus that we had previously used for our cross-linguistic study of context effects on consonant voicing (Shih and Möbius, 1998; Shih et al., 1999). The latter corpus is a subset (male speaker k61; 598 sentences, 12092 consonants, 4303 stops) of the Kiel Corpus of Read Speech (Kiel Corpus, 1994), which has manually labeled phone boundaries, including boundaries between the closure and release phases of stop consonants. The evident similarity of the stop consonant voicing profiles in the two corpora (Figure 14) increases the confidence in the automatic aligner's performance. It does not seem to be the case that analysis frames containing quasi-periodic energy in the transitional area between a vocalic or sonorant sound and a following plosive are erroneously assigned to the closure phase by the aligner; in fact, more voiced frames during the closure phase of stops are observed in the manually labeled corpus k61.

Glottal stops have not yet been analyzed with respect to their voicing properties. The aligner assigns the label "glottal stop" to any segment of speech that occurs at the appropriate place in the phone sequence (e.g., at vowel-vowel transitions across phonological word boundaries) and that matches its pertinent acoustic model. Phenomenologically, both local (glottalization, glottal stop) and distributed laryngealizations (creaky voice, creak, vocal fry, diplophonic phonation) (Hedelin and Huber, 1990) occur in the corpus and may be marked by the aligner as a glottal stop. We would be interested in the voicing properties of true glottal stops only, but without a manual labeling of the corresponding speech events there is no way of distinguishing the different types of laryngealizations in the corpus.

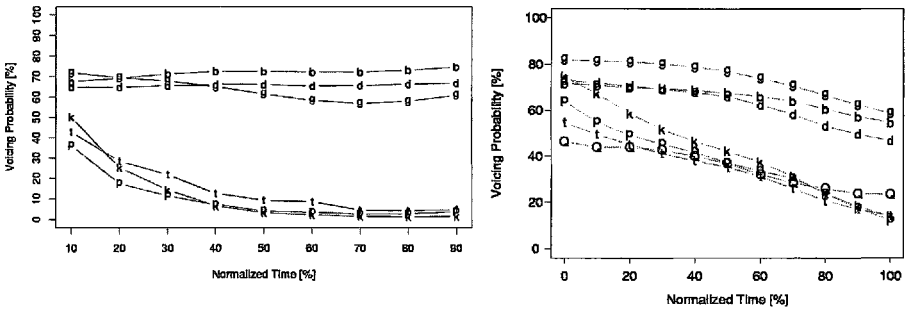

Figure 14: Voicing profiles of German stop consonant closures: MS corpus (left), k61 corpus (right, adapted from Shih and Möbius, 1998).

Vocal fold vibration is a fragile and complex process; it tends to break down unless specific conditions are satisfied. Several articulatory gestures are known to support voicing: (a) adducing the vocal folds; (b) slackening the vocal folds; (c) generating sufficient subglottal pressure; (d) producing a sufficient degree of mouth opening; (e) enlarging the vocal cavity. Each of these articulatory goals contributes to the maintenance of voicing; usually voicing is maintained by a combination of some of these gestures, and several trading relations exist between the individual gestures (Jessen, 2001).

The articulatory goal of enlarging the vocal cavity can in turn be achieved by a number of second-level variable articulatory goals, which may be described as follows: (e1) raising the velum; (e2) lowering the velum; (e3) advancing the tongue root; (e4) lowering the jaw; (e5) lowering the larynx. Again, trading relations exist between several of these goals, and some of them are in fact mutually exclusive, e.g. (e1) and (e2).

The analysis of voicing in German fricatives (section 3.2) suggests that there is a negative correlation between the absolute duration of the speech sound and its voicing probability. The constriction of the vocal tract required to produce a (voiced) fricative causes difficult aerodynamic conditions for vocal fold vibration; the longer the constriction is maintained the more likely voicing is to break down.

One should expect this effect to be even stronger in stop consonants (Kingston and Diehl, 1994), where the closure of the vocal tract prevents the maintenance of appropriate sublaryngeal and supralaryngeal pressure ratios. However, the analysis of stop consonant voicing profiles (section 3.1) revealed only a weak effect of a lower voicing probability in the later region of the closure phase of phonemically voiced stops (Figure 5), whereas a direct correlation of voicing with absolute closure duration was not found. In contrast, the expected significant effect was indeed found in Mandarin Chinese stop consonants (Shih et al., 1999).

A weak effect of place of articulation can be observed in Figure 5, where the probability of closure voicing is lower for [g] than for [b] and [d]. Given the

fact that enlargement of the vocal cavity supports the maintenance of voicing, a closure formed relatively far back in the vocal tract, as required for the production of velar stops, should in fact be expected to reduce the voicing probability of [g]. Size of vocal cavity also conceivably contributes to the higher voicing probability found in [v] than in [z] and [ʒ] (Figure 7).

This study has been motivated by the intention that phonological specifications of speech sounds be informed by phonetic data analysis, and that methodologies like the voicing profile approach can serve this purpose. The present approach and its results also have implications for applications in speech technology, in particular speech synthesis, automatic speech recognition, and automatic speech segmentation. Such applications are often sensitive to the discrepancies between the assumed specification of a speech sound and its acoustic realization. The mapping of the feature [voice] to the gradient, dynamically changing acoustic voicing status is problematic; the difficulty may be alleviated by means of probabilistic models that facilitate the context-sensitive prediction of voicing and voicing profiles. Such models based on the voicing profile method have been presented elsewhere (Shih and Möbius, 1998; Shih et al., 1999).

Note

* Acknowledgments: The author gratefully acknowledges the contributions by Bhuvana Narasimhan and Chilin Shih to the cross-linguistic part of this study. I thank Greg Dogil and especially Michael Jessen for insightful comments and suggestions. I also thank audiences at the Universities of Frankfurt, Köln and Stuttgart for their feedback.

Address of the Author:

Bernd Möbius
Institut für Maschinelle Sprachverarbeitung
Experimentelle Phonetik
Universität Stuttgart
Azenbergstraße 12
70174 Stuttgart, GERMANY
bernd.moebius@ims.uni-stuttgart.de

Bibliography

Chomsky, Noam & Morris Halle. 1968. The Sound Pattern of English. New York: Harper & Row.

Dixit, R. Prakash & W. S. Brown. 1985. "Peak Magnitudes of Oral Air Flow During Hindi Stops (Plosives and Affricates)." Journal of Phonetics 13, 219-234.

Halle, Morris & Kenneth N. Stevens. 1971. "A Note on Laryngeal Features." MIT Research Laboratory of Electronics Quarterly Progress Report 101, 198-213.

Hedelin, Per & Dieter Huber. 1990. "Pitch Period Determination of Aperiodic Speech Signals." Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing, 361-364.

IMS German Festival Home Page. 2003. [http://www.cstr.ed.ac.uk/projects/festival/].

Jakobson, Roman & Morris Halle. 1968. "Phonology in Relation to Phonetics." In: Bertil Malmberg (ed.), Manual of Phonetics. Amsterdam: North Holland, 411-449.

Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. Preliminaries to Speech Analysis. Cambridge, MA: MIT Press.

Jessen, Michael. 1999. "Redundant Aspiration in German is Primarily Controlled by Closure Duration." Proceedings of the 14th International Congress of Phonetic Sciences (San Francisco), Vol. 2, 993-996.

Jessen, Michael. 2001. "Phonetic Implementation of the Distinctive Auditory Features [voice] and [tense] in Stop Consonants." In: Tracy A. Hall & Ursula Kleinhenz (eds.), Recent Developments in Distinctive Feature Theory. Berlin: Mouton de Gruyter.

Jessen, Michael & Catherine Ringen. 2002. "Laryngeal Features in German." Phonology 19, 1-30.

Keating, Patricia. 1984. "Phonetic and Phonological Representation of Stop Consonant Voicing." Language 60, 286-319.

Kiel Corpus. 1994. The Kiel Corpus of Read Speech, Vol. 1. Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel. CDROM.

Kingston, John & Randy L. Diehl. 1994. "Phonetic Knowledge." Language 70, 419-455.

Kohler, Klaus J.. 1995. Einführung in die Phonetik des Deutschen. Berlin: Erich Schmidt Verlag. 2nd edition.

Ladefoged, Peter & Ian Maddieson. 1996. The Sounds of the World's Languages. Oxford: Blackwell.

Lisker, Leigh & Arthur S. Abramson. 1964. "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements." Word 20, 384-422.

Niyogi, Partha & Padma Ramesh. 2003. "The Voicing Feature for Stop Consonants: Recognition Experiments with Continuously Spoken Alphabets." Speech Communication. In press.

Pompino-Marschall, Bernd. 1995. Einführung in die Phonetik. Berlin: de Gruyter.

Poon, Pamela G. & Catherine A. Mateer. 1985. "A Study of VOT in Nepali Stop Consonants." Phonetica 42, 39-47.

R-Project. 2001. The R Project for Statistical Computing. [http://www.R-project.org/].

Rapp, Stefan. 1995. "Automatic Phonemic Transcription and Linguistic Annotation from Known Text with Hidden Markov Models: An Aligner for German." Proceedings of ELSNET Goes East and IMACS Workshop "Integration of Language and Speech in Academia and Industry" (Moscow, Russia).

Schweitzer, Antje, Norbert Braunschweiler, Tanja Klankert, Bernd Möbius & Bettina Säuberlich. 2003. „Restricted Unlimited Domain Synthesis." Proceedings of the European Conference on Speech Communication and Technology (Geneva), 1321-1324.

Shih, Chilin & Bernd Möbius. 1998. "Contextual Effects on Voicing Profiles of German and Mandarin Consonants." Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia), 81-86.

Shih, Chilin, Bernd Möbius & Bhuvana Narasimhan. 1999. "Contextual Effects on Consonant Voicing Profiles: A Cross-Linguistic Study." Proceedings of the 14th International Congress of Phonetic Sciences (San Francisco), Vol. 2, 989-992.

Shimizu, Katsumasa. 1990. A Cross-Language Study of Voicing Contrasts of Stop Consonants in Asian Languages. PhD Thesis, University of Edinburgh.

SmartKom. 2003. Das Leitprojekt SmartKom: Dialogische Mensch-Technik-Interaktion durch koordinierte Analyse und Generierung multipler Modalitäten. [http://smartkom.dfki.de/start.html].

Stock, Dieter. 1971. Untersuchungen zur Stimmhaftigkeit hochdeutscher Phonemrealisationen. Hamburg: Buske.

Wiese, Richard. 2000. The Phonology of German. Oxford: Oxford University Press.