

CONTEXTUAL EFFECTS ON VOICING PROFILES OF GERMAN AND MANDARIN CONSONANTS

Chilin Shih, Bernd Möbius

Bell Labs – Lucent Technologies, Murray Hill, NJ 07974, USA

ABSTRACT

In this paper we present a study of the voicing profiles of consonants in Mandarin Chinese and German. The voicing profile is defined as the frame-by-frame voicing status of a speech sound in continuous speech. We are particularly interested in discrepancies between the phonological voicing status of a speech sound and its actual phonetic realization in connected speech. We further examine the contextual factors that cause voicing variations and test the cross-language validity of these factors. The result can be used to improve speech synthesis, and to refine phone models to enhance the performance of automatic speech segmentation and recognition.

1. INTRODUCTION

There is ample documentation in the literature showing that the realization of voicing in many languages is highly variable. The frequent mismatch between phonological specification and acoustic realization in speech leads to complications in speech processing, notably in speech synthesis and speech recognition. Our main focus in this paper is to address implications for speech synthesis.

Previous studies investigating voicing concentrated on a few broad areas: voice onset time (VOT), perception, production, and contextual effects. It was noted, for instance, that the phonologically voiced stops in English do not necessarily have voicing throughout the closure duration, particularly in sentence initial and final position, an observation that even called into question the validity of voicing as a distinctive feature in English. It was suggested (e.g., [8, 9]) that VOT as a single-dimension continuous variable is capable of capturing 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.

A number of other acoustic cues, in particular duration, fundamental frequency (F_0), and formant trajectories, has been shown to enable listeners to make the voiced/voiceless distinction, provided that listeners can reliably perform the task without lexical information. Compared to its voiceless counterpart in otherwise identical context, a voiced consonant typically has shorter duration [1], with the preceding vowel having longer duration [2]. Voicing of a consonant has a lowering effect on F_1 [18, 17] and on F_0 [5, 20, 21] of surrounding vowels. In contrast, a voiceless consonant raises F_0 in the onset of a following vowel [7, 11]. It was also found that the voicing status of a consonant has an effect on tongue body kinematics [10]. A physiological account of the production of voiced and voiceless fricatives that explains some of the mentioned effects was provided in [17].

Effects of voicing have already been handled in different components in the Bell Labs text-to-speech (TTS) system [16]. For instance, the duration module predicts shorter durations of voiced consonants mostly in terms of intrinsic duration, while vowel length distinctions are modeled primarily by using the following phone as a factor. The intonation module provides segmental F_0 perturbation curves. F_1 transitions and formant changes as a result of tongue body variations are kept intact in the concatenative unit inventory.

However, there are several problems we are facing in TTS with regard to voicing. Our current unit concatenation algorithm assumes a constant status of voicing throughout the duration of a given phone. We feel that we would have to model natural variations of voicing within a phone to achieve higher naturalness. This has implications for the acoustic inventory construction procedure. Acoustic unit selection will have to pay special attention to voicing variation in natural speech, and therefore in candidate units. A unit that starts or ends with an unexpected voicing property will, if undetected, lead to unsmooth concatenation between units. Furthermore, in automatic speech segmentation, a discrepancy in voicing status between speech and transcription is likely to yield suboptimal segmentor performance.

In order to address these TTS specific problems in the future, we first need information on the contextual factors that affect voicing in speech. Some of the factors leading to voicing variation in speech were explored in [3, 15]. Our study extends the scope to all consonants in two languages, Mandarin Chinese and German, with speech databases that contain rich combinations of phones and contextual factors. In Section 3 we present an analysis of *voicing profiles* that show voicing variations over time within a phone, and the contrasts among phones and phone classes. The voicing profile is defined as the frame-by-frame voicing status of a speech sound in continuous speech. In Section 4 we present statistical models that predict voicing probability from phone identity, neighbouring phones, and positional and prosodic factors.

2. SPEECH DATABASE

The Mandarin database consists of 424 short paragraphs read by a male Beijing Mandarin speaker; it had previously been prepared for a study on segmental duration [14]. The database was selected using a greedy algorithm to maximize the occurrence of Mandarin phones in combination with preceding and following phones, tones, prosodic contexts, and syllable structures. For this paper 12,991 syllable-initial consonants were analyzed. The German database is a subset of the Kiel Corpus of Read Speech [4]. The subset consists of 598 sentences, read by a male speaker. A total of 12,092 consonants was analyzed. Voicing information was obtained automatically using the ESPS tools of Entropic Inc.

	lab	alv	pal	retro	vel	glot
vl asp stop	p	t			k	
vl unasp stop	b	d			g	
vl asp affr		c	j	C		
vl unasp affr		z	q	Z		
vl fric	f	s	x	S		h
vd fric	v			r		
nasal	m	n			N	
lateral		l				

Table 1: Mandarin consonant inventory.

3. VOICING PROFILES

This section presents the voicing profiles of Mandarin and German consonants. The voicing profile plots show the dynamic changes of voicing probability of a given speech sound as a function of normalized time. We obtained 11 voicing samples from each consonant in the database at 10 equidistant time intervals, including the initial and final frames. The voicing probability of a sound at a given position is calculated as the percentage of the population that is voiced at that position. In the voicing profiles, voicing probability is plotted on the y-axis as a function of 11 normalized time points, shown as percentage of the consonant duration. For example, Figure 1 shows that 83% of realizations of the voiceless unaspirated alveolar stop /d/ in Mandarin are voiced at the beginning of the closure, and 50% are still voiced at the end of the closure.

The Mandarin transcription convention, as given in Table 1, follows fairly closely the standard romanization system *pinyin*. However, all digraphs in pinyin are now represented by single letters: the retroflex consonants /zh, ch, sh/ in pinyin are transcribed as /Z, C, S/, and the velar nasal /ng/ is transcribed as /N/. For the sake of consistency, we use the same set of symbols for the corresponding sounds in German, where appropriate (Table 2).

The voicing profiles of the closure phases of Mandarin stops are shown in Figure 1. The most robust effect is the clear separation of unaspirated and aspirated sounds. Unaspirated stops and affricates (not shown) are more likely to be voiced than aspirated ones. This result is consistent with the finding that the vocal cords come closely together to build up pressure before the release of aspirated consonants, and hence interrupt voicing momentarily. The

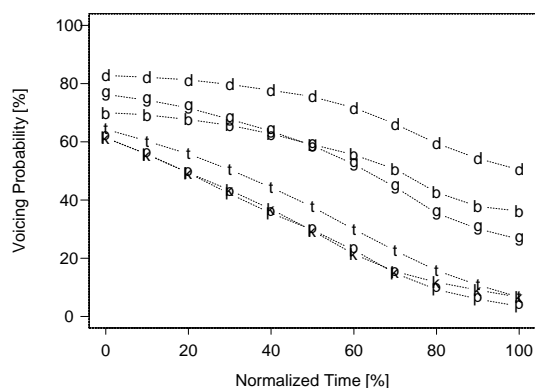


Figure 1: Voicing profiles of Mandarin stops.

	lab	alv	alvpal	pal	vel	glot
vl stop	p	t			k	Q
vd stop	b	d			g	
vl fric	f	s	S	c	X	h
vd fric	v	z	Z			
nasal	m	n			N	
liq/gli		l		j	r	

Table 2: German consonant inventory.

German data display a remarkably similar picture (Figure 2): the populations of /b,d,g/ are neatly separated from those of /p,t,k/, at least after the initial 20% of the closure duration.

Phonologically, all Mandarin stop consonants are voiceless; /p,t,k/ are aspirated, /b,d,g/ are unaspirated. In German, stops are usually classified by voicing: /p,t,k/ are voiceless, /b,d,g/ are voiced; aspiration is an optional feature of the voiceless stops. The results of our study show that, despite this typological difference between the consonantal systems of the two languages, Mandarin /b,d,g/, phonologically voiceless and unaspirated, and German /b,d,g/, phonologically voiced and unaspirated, show very similar patterns in their voicing profiles.

Voicing profiles of fricatives and sonorants are given in Figure 3 for Mandarin and Figure 4 for German. The Mandarin data are separated into two distinct populations, the voiced consonants /l,r,m,n,v/ and the voiceless fricatives /f,s,S,x,h/. The voiced consonants typically show voicing throughout. There are a few voiceless initial frames, all of which occur in sentence initial position. A considerable proportion of voiceless fricatives show voicing at the two edges of the consonant. Fricatives in Mandarin must occur in either sentence initial or intervocalic position. The high percentage of voicing at the edges of fricatives is clearly due to the influence of adjacent voiced sounds. Among phonologically voiceless fricatives, /h/ is most susceptible to voicing (see also [13]). 20% of the /h/ population show voicing throughout their duration. Excluding /h/, all other voiceless fricatives maintain voiceless frames at least near the mid point of the consonant.

In the German data a three-way separation of sounds is found. Nasals, liquids and glides form a consistent group that is voiced for most of the consonant duration, but shows a significant de-voicing effect in the onset phase. The second group consists of the

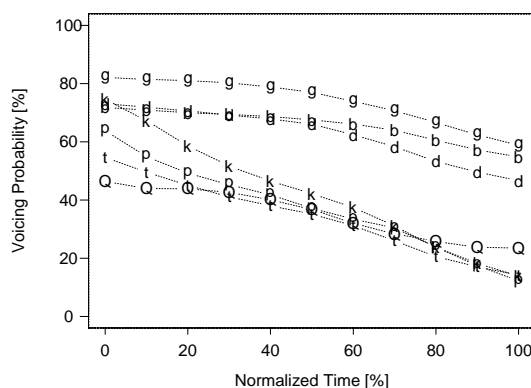


Figure 2: Voicing profiles of German stops.

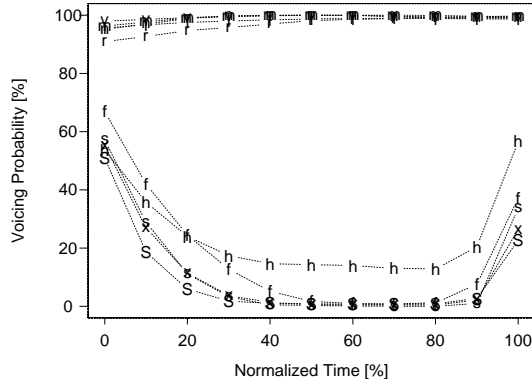


Figure 3: Voicing profiles of Mandarin sonorants and fricatives.

fricatives /f,s,S,c/, which often start voiced but are typically voiceless by the mid point, and of /x,z,Z/, which have almost identical profiles, but with a higher percentage of voiced frames. The third group is comprised of /v,h,r/, whose voicing profiles are entirely predictable from segmental context.

Considerable differences can be observed in the realization of sonorants in the two languages. In Mandarin, /m,n,l,r/ are solidly voiced throughout; only in isolated cases are initial frames voiceless. German /m,n,N,l,j/ are often initially unvoiced, conditioned by preceding voiceless obstruents. These findings reflect differences in the syllable structure of the two languages. Mandarin does not allow consonant clusters, and all syllable final consonants must be nasal. Consequently, all sentence medial consonants in Mandarin occur in a voiced environment, so phonologically voiced consonants tend to maintain voicing throughout the consonant duration. In German, consonant clusters are common in syllable onsets and codas. The voicing properties of surrounding phones are more varied and have an immediate impact on conditioning the voicing probability of a consonant.

4. VOICING MODELS

This section discusses statistical models that predict the probability of voicing from phone identity, neighbouring phones, and positional and prosodic factors.

The speech databases of Mandarin and German were coded for eight and seven factors, respectively, to train additive models that

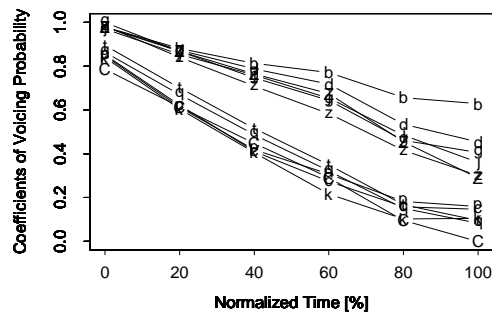


Figure 5: Coefficients of voicing prediction models in Mandarin: identities of stops and affricates.

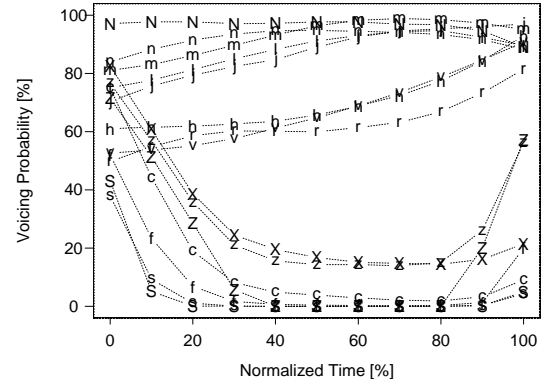


Figure 4: Voicing profiles of German sonorants and fricatives.

predict voicing; see [19] for mathematical proof and calculation. Voicing probability is represented in the input data as a number ranging between 0 and 1, with 0 indicating voicelessness and 1 indicating voicedness. From each consonant 11 samples were taken at equidistant time intervals, and additive models were trained for each of these 11 positions. For voicing prediction, coefficients of the relevant levels from all factors were summed up. The voicing threshold was set at 0.5.

4.1. Mandarin Chinese

In this section voicing probability models for Mandarin are presented. The discussion concentrates on stops and affricates, which show the most interesting effects. The following factors were used; the number of levels on each factor is indicated in parentheses: 1. phone identity (12: b, p, ...); 2. phone duration (10: 10–100 ms); 3. preceding tone (4); 4. preceding phone (9 classes); 5. tone of the current syllable (3); 6. following phone (8 classes); 7. prominence level (2); 8. distance to word initial position (3). Only the effects of factors 1, 2, 5 and 8 will be discussed here. The effects of the other factors are comparatively weak and show little variation throughout closure duration.

Figure 5 illustrates the effect of the most important factor, phone identity. All sounds show a decline of voicing probability as time proceeds. The aspirated and unaspirated populations are clearly separated. Aspirated sounds start voiced and become increasingly

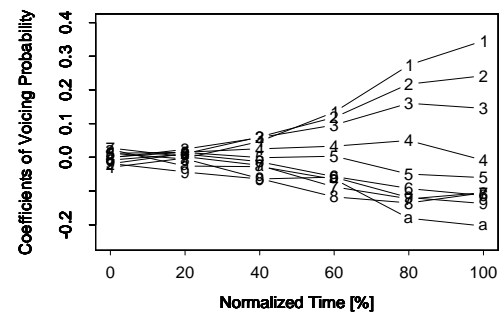


Figure 6: Coefficients of voicing prediction models in Mandarin: duration of stop closure. “1”–“9” = closure durations of 10–90 ms, “a” = longer than 90 ms.

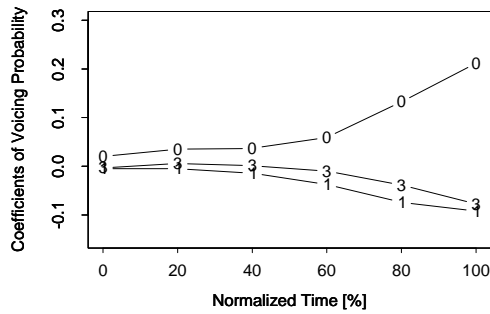


Figure 7: Coefficients of voicing prediction models in Mandarin: tones. 0 = neutral tone / unstressed syllables; 1 = tones that end high; 2 = tones that end low; 3 = tones that end low.

voiceless toward the end of the consonant. Within each population, stops are more likely to be voiced than affricates. There is an apparent discrepancy in that /d/ appears to be most likely to be voiced among all stops in Figure 1, but /b/ takes over in Figure 5. It turns out that many of the voiced samples of /d/ come from the reduced syllable *de*, the possessive, adjectival, and relative clause marker. The coefficients shown in Figure 5 were corrected for the reduction effect that is expressed by both shorter duration and the neutral tone.

The effect of closure duration on voicing is displayed in Figure 6. Closure duration has no effect on the first half of the consonant but the effect becomes increasingly strong toward the end of the closure. The magnitude of the coefficients is completely in line with the duration: shorter duration correlates with higher voicing probability.

Figure 7 illustrates the effect of tone. Consonants in a full tone syllable, represented as 1 (tones that end high) or 3 (tones that end low), are less likely to be voiced. Consonants in neutral tone syllables, represented as 0, which are comparable to unstressed syllables in English, are more likely to be voiced in the second half of the consonant. As with the duration coefficients, the effect becomes stronger toward the end of the consonant.

The effect of the position of the syllable in the word is illustrated in Figure 8. The symbols 0, 1 and 2 represent word initial, second syllable, and third or later syllable in the word, respectively. Word initial stops are least likely to be voiced, and stops in the second

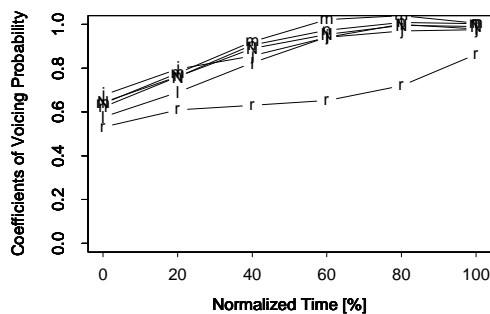


Figure 9: Coefficients of voicing prediction models in German: sonorants in syllable onsets.

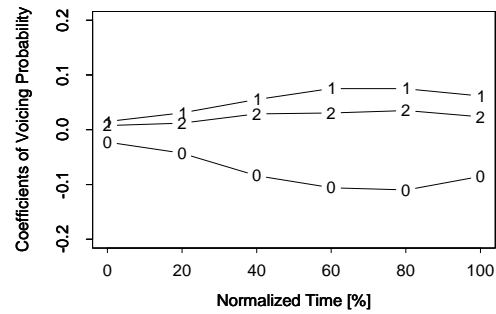


Figure 8: Coefficients of voicing prediction models in Mandarin: syllable position in the word. Numbers 0, 1, and 2 represent the distance from the beginning of the word, 0 being word initial.

syllable are most likely to be voiced. The effect, again, tends to be stronger near the end of the consonant.

Many factors that enhance the probability of voicing in Mandarin reflect prosodically weak positions, which cause lenition, and therefore higher voicing probability, of the phonologically voiceless consonants. These effects include shorter duration, discourse prominence on the preceding syllable, neutral tone (de-stressing) of the current syllable, and word-medial position.

4.2. German

In our study of the voicing profiles of German consonants the following factors were used; the number of levels on each factor is indicated in parentheses: 1. phone identity (15: m, n, v, ...); 2. word class (2: content word, function word); 3. stress (2); 4. location in syllable (2: onset, coda); 5. preceding phone (6 classes); 6. following phone (6 classes); 7. phone duration (10: 10–100 ms). Only the effects of factors 1, 4, and 5 will be discussed here. The effects of the other factors are comparatively weak and show little systematic variation.

The most interesting effects can be observed in sonorants and phonologically voiced fricatives and /h/. Figure 9 displays the effect of the phone identity of nasals, liquids and glides in syllable onset position, including all preceding segmental contexts (the following context is almost always a vowel). All these sounds show a considerable degree of devoicing in the first half of their total durations. Whereas /r/ is often devoiced throughout, most

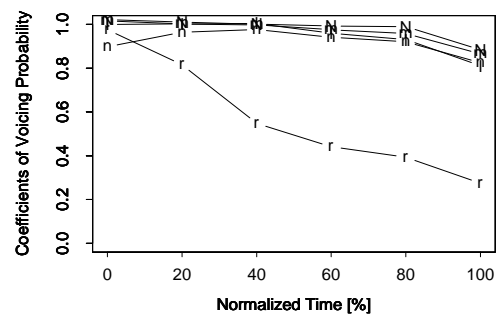


Figure 10: Coefficients of voicing prediction models in German: sonorants in syllable codas.

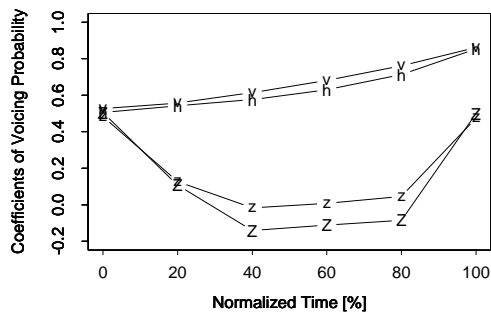


Figure 11: Coefficients of voicing prediction models in German: phonologically voiced fricatives and /h/ in syllable onsets.

sonorants become solidly voiced in the second half of their durations.

In syllable coda position, no devoicing of sonorants is generally observed, with the exception of /r/, which has more than 50% voiceless frames in the second half of its duration (Figure 10). This observation can be explained by distributional properties of this speech sound: in German syllable codas, /r/ can only occur as a morpheme-final consonant, e.g. in *starr* [Star], and it may then neutralize the voicing opposition as it is the case for fricatives and stops. Otherwise, /r/ is vocalized if it is followed by another consonant in the same syllable coda, e.g. *hart* [hært].

The voicing probability models for phonologically voiced fricatives and /h/ in syllable onsets are given in Figure 11, averaged across preceding contexts. While /v/ and /h/ follow the pattern observed for /r/ in the same location (Figure 9), /z/ and /Z/ are solidly unvoiced in the center in any context. In fact, the probability of /Z/ being voiced in the center is so low that the coefficients of the additive model have negative values.

The single most important factor determining the voicing of German consonants is the immediate segmental context. Figure 12 displays the voicing probability models for sonorants, phonologically voiced fricatives and /h/ in syllable onset position and preceded by voiceless obstruents. A strong devoicing effect is observed on all these speech sounds. The effect is significant for all phones throughout the first half of their durations, and all phones other than /m,n,l,j/ are affected in their entire durations. As noted before, /z/ is completely voiceless.

In the context of preceding sonorants and vowels, most sonorants, phonologically voiced fricatives and /h/ can be reliably predicted as voiced in syllable onsets (Figure 13). Exceptions are again the fricatives /z/ and /Z/, which are typically voiceless in the center, even in intervocalic position.

The observation that the devoicing effect is strongest on /z/ and /Z/, which are typically devoiced at least in the center, disregarding the segmental context, can be partially explained as follows. First, German phonotactics does not allow a syllable onset to consist of the voiceless fricative /s/; therefore, no minimal pairs exist in this position, and a complete devoicing of /z/ even in intervocalic context does not lead to a confusion of the two sounds. Second, a near merger between /Z/ and /S/ is facilitated by the

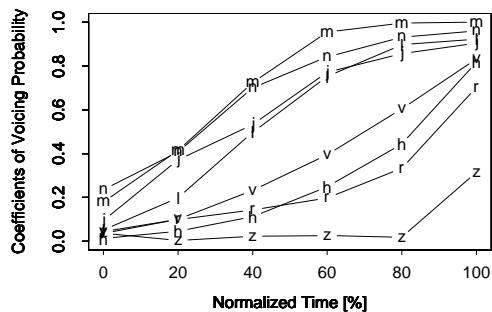


Figure 12: Coefficients of voicing prediction models in German: effect of preceding voiceless obstruents on sonorants, phonologically voiced fricatives and /h/ in syllable onsets.

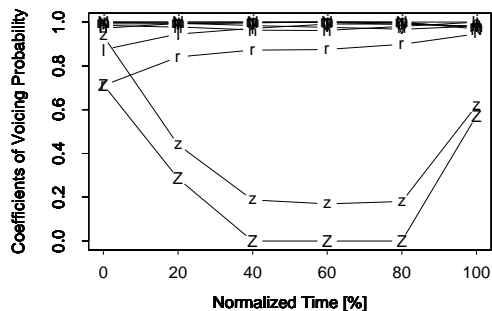


Figure 13: Coefficients of voicing prediction models in German: effect of preceding vowels and sonorants on sonorants, phonologically voiced fricatives and /h/ in syllable onsets.

fact that /Z/ only occurs in foreign loan words; minimal pairs are virtually non-existent.

Devoicing of phonologically voiced fricatives and /j/ does not lead to a perceptual merger of /v/ and /f/, /z/ and /s/, or /j/ and /c/. The overall frication energy is considerably higher in the phonologically voiceless fricatives.

Nasal consonants and /l/ have no voiceless counterparts in the phonemic system of German. As for /r/, a confusion with the voiceless fricative /X/ is possible but somewhat restricted due to the distributional properties of the two sounds.

It is remarkable that the presence of a syllable boundary between an unvoiced obstruent and the following sonorant or voiced fricative does not significantly reduce the devoicing effect (see also [6, p. 158]).

5. CONCLUSION

In this paper we presented a study of the voicing profiles of consonants in Mandarin Chinese and German. The voicing profile is defined as the frame-by-frame voicing status of a speech sound in continuous speech. The voicing probability of a sound at a given position was calculated as the percentage of the population that is voiced at that position. We discussed discrepancies between the phonological voicing status of a speech sound and its actual phonetic realization in connected speech. We further examined

the contextual factors that cause voicing variations and tested the cross-language validity of these factors.

Statistical models were proposed that predict the probability of voicing in Mandarin and German from phone identity, neighbouring phones, and positional and prosodic factors. A speech database was coded for these factors. Samples were taken from each consonant at equidistant time intervals, and additive models were trained for each of these positions. For voicing prediction, coefficients of the relevant levels from all factors were summed up.

The results of this study have potential impact on both speech synthesis and recognition. The voicing variation of a speech sound in a given context can be modeled directly in a formant synthesizer. However, in the case of a concatenative TTS system, such as the Bell Labs system, issues in inventory design are affected. Context-dependencies and coarticulatory effects are the main obstacles to a straightforward computation of which units are needed in the acoustic inventory of a given language. For instance, these effects can require the use of inventory elements whose domain is larger than that of a diphone.

To incorporate our findings into the unit selection procedure [12], the voicing profile has to be used as an explicit criterion, in addition to commonly applied criteria such as spectral distance measures. Moreover, finer control of the source model needs to be implemented to generate the desired voicing variation within a phone. Predictable voicing variation may also be exploited in context-sensitive phone modeling in speech recognition, and in automatic speech segmentation.

6. REFERENCES

1. Crystal, T. H., and House, A. S. Segmental durations in connected speech signals: current results. *Journal of the Acoustical Society of America* 83 1988, 1553–1573.
2. Denes, P. Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America* 27 1955, 761–764.
3. Haggard, M. The devoicing of voiced fricatives. *Journal of Phonetics* 6 1978, 95–102.
4. The Kiel Corpus of Read Speech, vol. 1. Published by Institut für Phonetik und Digitale Sprachverarbeitung, Universität Kiel, Germany, 1994. CDROM.
5. Kohler, K. J. F_0 in the perception of lenis and fortis plosives. *Journal of the Acoustical Society of America* 78 1985, 21–39.
6. Kohler, K. J. *Einführung in die Phonetik des Deutschen*, 2nd ed. Erich Schmidt Verlag, Berlin, 1995.
7. Lehiste, I., and Peterson, G. Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America* 33 1961, 419–425.
8. Liberman, A., Delattre, P., and Cooper, F. The role of selected stimulus variables in the perception of voiced and voiceless stops in initial position. *Language and Speech* 1 1958, 153–167.
9. Lisker, L., and Abramson, A. S. A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20 1964, 384–422.
10. Löfqvist, A., and Gracco, V. L. Tongue body kinematics in velar stop production: influences of consonant voicing and vowel context. *Phonetica* 51 1994, 52–67.
11. Mohr, B. Intrinsic variations in the speech signal. *Phonetica* 23 1971, 65–93.
12. Olive, J., van Santen, J., Möbius, B., and Shih, C. Synthesis. In Sproat [16], ch. 7, pp. 191–228.
13. Pierrehumbert, J. B., and Talkin, D. Lenition of /h/ and glottal stop. In *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, G. J. Docherty and D. R. Ladd, Eds. Cambridge University Press, 1992, pp. 90–115.
14. Shih, C., and Ao, B. Duration study for the Bell Laboratories Mandarin text-to-speech system. In *Progress in Speech Synthesis*, J. van Santen, Ed. Springer, New York, 1997, pp. 383–399.
15. Smith, C. The devoicing of /z/ in American English: effects of local and prosodic context. *J. Phon.* 25 1997, 471–500.
16. Sproat, R. W., Ed. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer, Dordrecht, 1998.
17. Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowski, K. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America* 91 1992, 2979–3000.
18. Stevens, K. N., and Klatt, D. H. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America* 55 1974, 653–659.
19. van Santen, J. P. H. Contextual effects on vowel duration. *Speech Communication* 11 1992, 513–546.
20. Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica* 47 1990, 36–49.
21. Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. F_0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America* 93, 4 1993, 2152–2159.