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., [6, 7]) 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 [15, 14] and on F_0 of surrounding vowels [5, 17]. In contrast, a voiceless consonant raises F_0 in the onset of a following vowel. It was also found that the voicing status of a consonant has an effect on tongue body kinematics [8]. A physiological account of the production of voiced and voiceless fricatives that explains some of the mentioned effects was provided in [14].

Effects of voicing have already been handled in different components in the Bell Labs text-to-speech (TTS) system [13]. 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, 12]. 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 [10]. 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., with some manual inspection and correction.

	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 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.

	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.

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 ara 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 [9]). 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. Sonorants, liquids and glides form a consistent group that is voiced for most of the consonant duration, but shows a significant devoicing effect in the onset phase. The second group consists

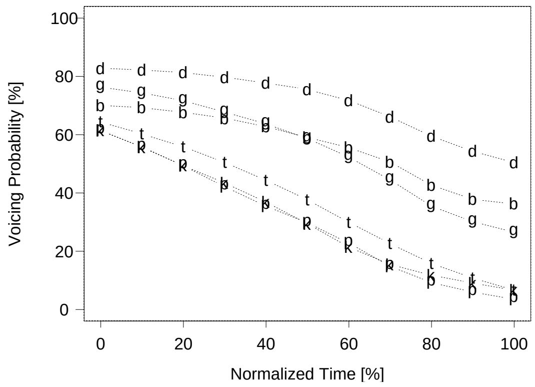

Figure 1: Voicing profiles of Mandarin stops.

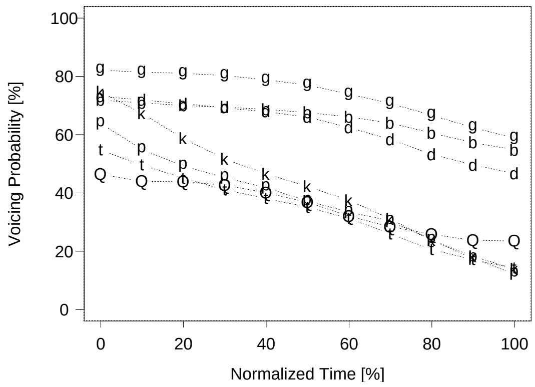

Figure 2: Voicing profiles of German stops.

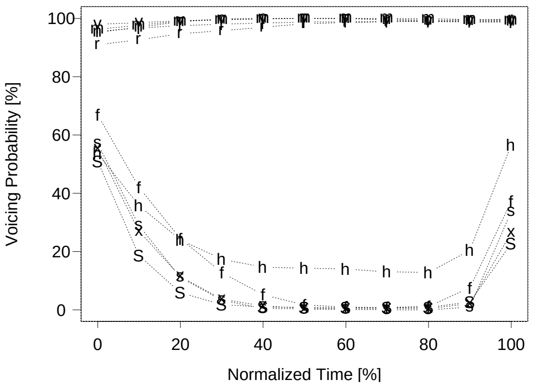

Figure 3: Voicing profiles of Mandarin sonorants and fricatives.

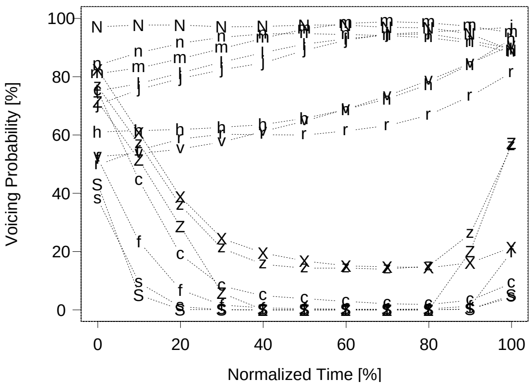

Figure 4: Voicing profiles of German sonorants and fricatives.

of the 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. Due to space constraints, only models for Mandarin stops and affricates will be presented. More details on the models for Mandarin and German are presented elsewhere [11]; the factorial schemes vary with the particular models. The single most important contextual factor in German is the immediate segmental context [11].

The database was coded for eight factors to train additive models that predict voicing in Mandarin; see [16] 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. 11 samples were taken from each consonant at equidistant time intervals, and 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.

For stops and affricates, 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, and 5 will be discussed here. The effects of the other factors are comparatively weak and show little variation throughout closure duration.

Figure 5 shows 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

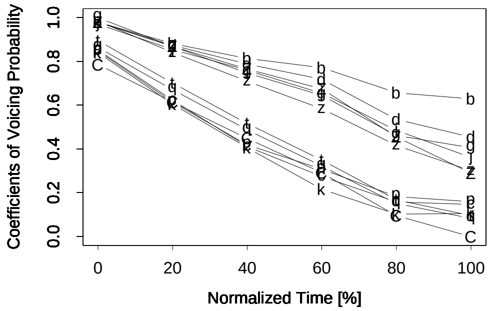

Figure 5: Coefficients of voicing prediction models: identities of stops and affricates.

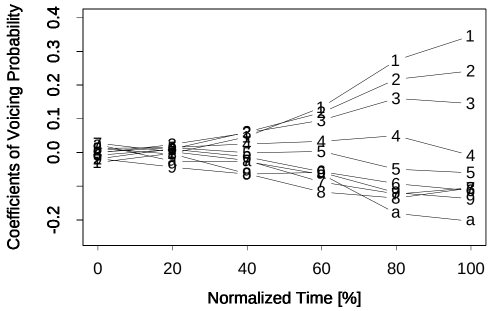

Figure 6: Coefficients of voicing prediction models: duration of stop closure. “1”–”9” = closure durations of 10–90 ms, “a” = longer than 90 ms.

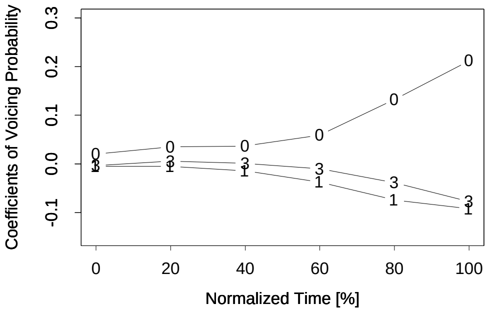

Figure 7: Coefficients of voicing prediction models: tones. 0 = neutral tone / unstressed syllables; 1 = tones that end high; 3 = tones that end low.

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.

Figure 6 shows the effect of closure duration on voicing. 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 shows 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.

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.

5. CONCLUSION

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, 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. *J. Acoust. Soc. Am. 83* 1988, 1553–1573.
2. Denes, P. Effect of duration on the perception of voicing. *J. Acoust. Soc. Am. 27* 1955, 761–764.
3. Haggard, M. The devoicing of voiced fricatives. *J. Phon. 6* 1978, 95–102.
4. The Kiel Corpus of Read Speech, vol. 1. Publ. by IPDS, Univ. Kiel, 1994.
5. Kohler, K. J. F_0 in the perception of lenis and fortis plosives. *J. Acoust. Soc. Am. 78* 1985, 21–39.
6. Liberman, A., Delattre, P., and Cooper, F. The role of selected stimulus variables in the perception of voiced and voiceless stops in initial position. *Lang. Speech 1* 1958, 153–167.
7. Lisker, L., and Abramson, A. S. A cross-language study of voicing in initial stops: acoustical measurements. *Word 20* 1964, 384–422.
8. 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.
9. 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.
10. 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, 1997, pp. 383–399.
11. Shih, C., and Möbius, B. Contextual effects on voicing profiles of German and Mandarin consonants. In *Proc. 3rd Int. Workshop Speech Synthesis (Jenolan Caves, Australia)* (1998).
12. Smith, C. The devoicing of /z/ in American English: effects of local and prosodic context. *J. Phon. 25* 1997, 471–500.
13. Sproat, R. W., Ed. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic Publishers, 1998.
14. Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowski, K. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *J. Acoust. Soc. Am. 91* 1992, 2979–3000.
15. Stevens, K. N., and Klatt, D. H. Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Am. 55* 1974, 653–659.
16. van Santen, J. P. H. Contextual effects on vowel duration. *Speech Comm. 11* 1992, 513–546.
17. 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.