Robustness of Tonal and Segmental Information in Noise - Auditory and Visual Contributions

Hansjörg Mixdorff, Yuping Wang ** and Yu Hu ***

*Faculty of Computer Science and Media, TFH Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
**University of Science and Technology, Hefei, China
ypwang|yuhu@iflytek.com

Abstract

This paper examines the robustness of tonal and segmental cues in noise exemplified by Mandarin monosyllables. We investigate how varying levels of noise inhibit the recognition of syllabic tone, onset, vowel nucleus and coda, and which property of the syllable is the most stable in audio only and audio plus video conditions.
A corpus of 220 frequent syllables was uttered by a male speaker of Mandarin and video-taped. Multi-talker babble noise was added to the resulting speech recordings at SNRs of 0, -3, -6 and -9 dB. In a perception test subjects were asked to write down the Pinyin plus tone combination of the word they perceived. Results indicate, inter alia, that the tonal information is more robust in noise than the segmental one, and that the nuclear vowel is the most stable part of the syllable. Auditory-visual gain was observed for the segments, but different from results of an earlier study, not for the tones. Tonal recognition rates were also influenced by the type of the nuclear vowel, with front vowels yielding highest and back vowels yielding lowest rates.

1. Introduction

Syllabic tones in tone languages are connected with distinct *F0* patterns (rising, falling etc.). Mandarin has four different lexical tones: high (1), rising (2), low (3), and falling (4), commonly used tone indices are given in brackets.

Research has shown that these tone contours patterns can be associated with distinct *F0* patterns. While this might suggest that tone is a purely acoustic phenomenon, there are now auditory-visual studies that suggest that speakers also exploit visual cues when identifying tones [1][2].

In an earlier study on Mandarin [3] we found that visual information enhances tone perception when the audio information is reduced by a masking noise. This auditory-visual gain rises as the SNR decreases. We did, however, not consider the effect of the noise on the segmental information. Therefore, strictly speaking, the so-called word identification test we performed was actually only a tone identification task as we provided the subjects with a selection of possible Chinese characters pertaining to the different choices. Subjects therefore knew the segmental content of the stimulus they had to classify. In the present study we aim to investigate whether tonal and segmental information is affected equally as the masking noise becomes stronger or whether the tonal information is more robust. Furthermore we want to examine the influence of the syllable's segmental structure on tonal identification, as well as auditory-visual gain.

2. Video material and stimulus materials

A corpus of 220 frequent Mandarin mono-syllabic words was uttered by a male speaker of Mandarin and video-taped with a mini DV camera in portrait mode (720 x 576 pixels, 25 frames per second, sampling rate 48 kHz) at TFH Berlin and transferred to a PC. The data were rotated by 90° and de-interlaced discarding even fields using *VirtualDub* [4]. The resulting video format (288 x 360 pixels, see Figure 1) was compressed using the Indeo 5.1 codec at maximum quality and the audio down-sampled to 16 kHz.

In order to segment the long video sequences into chunks of individual tokens, the audio tracks were annotated acoustically using *Praat TextGrid* [5] which were converted to *VirtualDub* scripts which in turn were used for automatically cutting the video as well as saving the associated soundtracks to individual wave files. The videos were cut with a window starting 400 ms before the acoustic onset of the syllable and ending 400 ms after the offset.

Babble noise was added to the speech signals at levels of 0, -3, -6 and -9 dB. This range was chosen after first informal tests showed that masking on audio only stimuli started at 0 dB, whereas stimuli were completely masked at -12 dB.

Figure 1: *Section of the speaker's face that was video-taped.*

The original audio tracks of the videos were then replaced by the noisy versions at 0, -3, -6 and -9dB SNR yielding

altogether eight types of stimuli which will be referred to as follows:
Noise0/3/6/9-A, babble-noise masked audio only stimuli
Noise0/3/6/9-AV, babble-noise masked audio plus video stimuli.

3. The Perception Test

Experiments were conducted using the *DMDX* software [6] employing scripts created by the first author. Considering the large number of stimuli and the fact that the tests were to be conducted with native Mandarin listeners, an identification task rather than a discrimination task was employed. The recognized phonetic content had to be specified using the romanized transcription system commonly known as Pinyin (see, for instance, [7]) including the number of the tone.

A designated set of tokens other than the 220 words was chosen for a practice session preceding the experiment proper, and the 220 words were divided into four groups. During a session each subject was presented with 220 stimuli from four different auditory or auditory-visual conditions in four consecutive blocks of trials. As can be seen in Table 1, a rolling design was employed such that the four types of stimuli presented to a particular participant in one trial set were, for instance, **Noise0-AV**, **Noise3-A**, **Noise6-AV**, and **Noise9-A,** with each block containing a different set of syllables, and the sequence of stimulus types varying between the four trial sets. Hence, the total number of 220 x 8 = 1760 stimuli was divided into eight different sets.

Table 1: *Structure of experiment trials with respect to stimulus type and syllable set.*

	trial block			
set	1	2	3	4
1	Words -1 Noise0-AV	Words-3 Noise3-A	Words-2 Noise6-AV	Words-4 Noise9-A
2	Words-2 Noise0-AV	Words-4 Noise3-A	Words-3 Noise6-AV	Words-1 Noise9-A
3	Words-3 Noise0-AV	Words-1 Noise3-A	Words-4 Noise6-AV	Words-2 Noise9-A
4	Words-4 Noise0-AV	Words-2 Noise3-A	Words-1 Noise6-AV	Words-3 Noise9-A
5	Words-1 Noise3-AV	Words-3 Noise0-A	Words-2 Noise9-AV	Words-4 Noise6-A
6	Words-2 Noise3-AV	Words-4 Noise0-A	Words-3 Noise9-AV	Words-1 Noise6-A
7	Words-3 Noise3-AV	Words-1 Noise0-A	Words-4 Noise9-AV	Words-2 Noise6-A
8	Words-4 Noise3-AV	Words-2 Noise0-A	Words-1 Noise9-AV	Words-3 Noise6-A

Within each block, tokens pertaining to a word set were presented consecutively, but in randomized order, and a session took less than 30 minutes to complete. Participants listened to the stimuli over headphones connected to a PC soundcard. Each trial started with a preparation phase of one and a half seconds during which the word 'ready' was displayed. Then the stimulus was presented, followed by the request to type the Pinyin sequence of the word perceived completed by hitting the <RETURN> key. Each stimulus was only presented once, and there was no time limit for the Pinyin input.
Participants were 16 members of staff (1 male, 15 had female) at iFlyTek corporation, Hefei, China, aged 21-33 from the speech data department. They reported to have normal hearing, four had corrected vision. All of them had received special training for speech annotation works. None of them was familiar with the speaker who had produced the video data. Each participant performed on one of the eight of the trial sets in Table 1. After the test, participants were interviewed about their observations. Six of the subjects felt that the test had been very tiresome, only five thought that the video had facilitated the task. A rather unexpected but important comment was that the speaker on the video was not handsome enough to deserve being looked at.
The data were aggregated, checked for typos or illegal inputs and subjected to statistical analysis.

4. Results

Figure 2 displays the proportion of correct responses for tones and syllable segments depending on the SNR. As can be seen, the rate drops steadily as the SNR decreases. Already at 0dB the syllable recognition is only 68% compared to 95% for the tones. At -9dB tone recognition approaches chance level, whereas the segments are only correctly identified in 6% of the times.

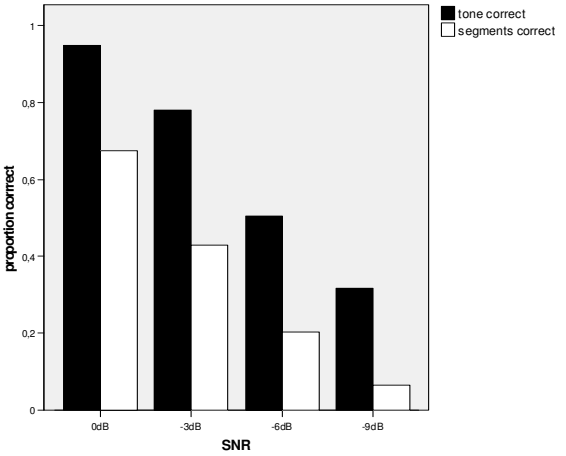

Figure 2: *Pooled results from perception experiment showing the proportion correct of segmental and tonal information depending on the SNR of the stimulus.*

Figure 3 displays the proportion correct for the four different tones of Mandarin. Just as shown in [3] the third tone is best identified in noise due to its relatively long syllabic duration and two-peaked intensity pattern. Tone 4 syllables in turn are relative short.

If we have a look at where in the syllable the recognition errors mostly occur, we yield the picture shown in Figure 4. As the SNR decreases, recognitions rates for all parts of the syllable, i.e. the onset, the nuclear vowel and the coda drop. In all cases but the lowest SNR of -9dB, rates are highest

for the nuclear vowel, followed by the onset and then the coda. Even at -9dB about 40% of at least one part of the syllable are identified correctly. At this SNR, onset-nucleus sequences are identified correctly in 23% of cases, and nucleus-coda sequences in 26%. It should be noted that after looking at some of the typical confusion patterns we treated diphthong nuclei such as 'ai' or 'ou' for this evaluation as sequences of nucleus 'a' and coda 'i', and nucleus 'o' and coda 'u', respectively.

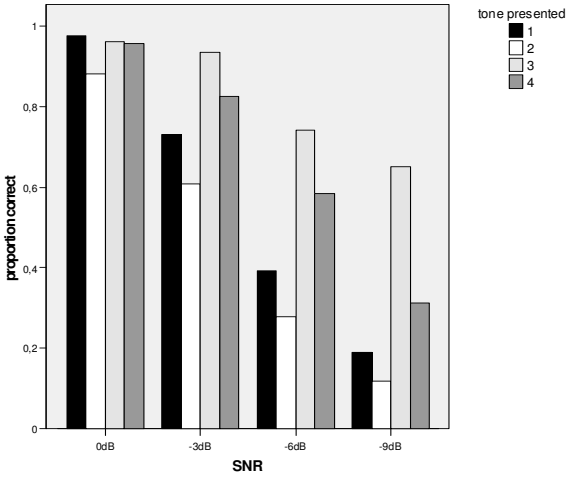

Figure 3: *Proportion correct for recognition of the four syllabic tones depending on the SNR of the stimulus. As can be seen tone 3 is the most stable.*

Figure 5 shows results concerning the potential auditory-visual gain yielded when the video image is presented along with the audio. The proportions correct for tones and segments depending on the SNR are displayed separately for audio only (video=0) and audio-visual (video=1) conditions. Different from the results found in [3] there is no tangible gain for the tonal discrimination whereas it is obvious for the segmental information. The syllable recognition rate rises by up to 20% with the visual information.

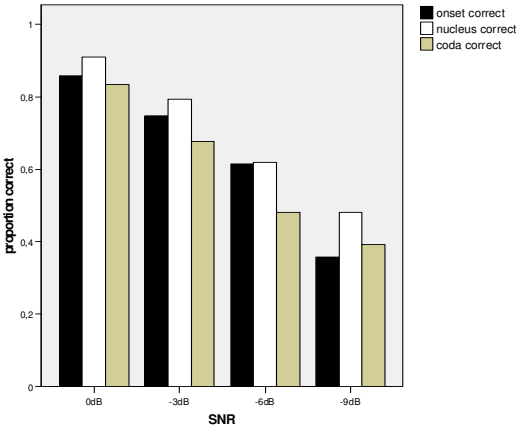

Figure 4: *Proportion correct for recognition of onset, nucleus and coda depending on the SNR of the stimulus. As can be seen, the nucleus is the most stable.*

We examined which syllables gained most from the visual information and looked at the proportion correct for different nuclear vowels. Results are displayed in Table 2. Since the Pinyin system is not always consistent with regards to vowel representation we chose SAMPA-like symbols for indicating the vowel property along with Pinyin examples. As could be expected, vowels with strong lip rounding such as [o], [y] and [u] benefit most from the visual information.

Table 2: *Proportion of correct nuclear vowel identification depending on the type of the nuclear vowel in audio only and audio plus video conditions.*

nuclear vowel	audio only	audio plus video	Pinyin example
@	,69	,75	sh**e**ng
a	,67	,84	sh**a**n
e	,75	,84	sh**ei**
E	,49	,76	ni**a**n
i	,70	,75	n**i**
I	,82	,79	sh**i**
o	,31	,70	p**o**
u	,43	,89	sh**u**
U	,24	,69	s**o**ng
y	,50	,88	y**u**
total	,61	,79	

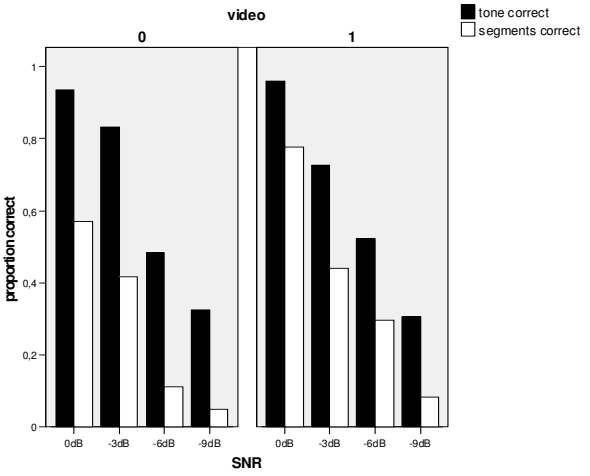

Figure 5: *Proportion correct for recognition of tonal and segmental information depending on the SNR of the stimulus, split into audio only (video=0, left) and audio plus video (video=1, right) conditions. The auditory-visual gain is significant in the recognition rate for the segments, but not for the tones.*

If we look at the relationship between the type of the onset consonant and the onset recognition rate, fricatives such as [s] and [S] get the highest rates, followed by plosives [p], whereas sonorants are affected most by the masking noise (see Table 3). In Table 4 we display a list of onsets prone to confusion and their most frequent confusion partners. The consonants are

represented using a SAMPA-type notation. As can be seen most of the confusions concern candidates that either share place or manner of articulation.

Table 3: *Proportion of correct onset identification depending on the type of initial consonant and SNR.*

	SNR			
vowel type	0dB	-3dB	-6dB	-9dB
fricatives	.90	.82	.74	.52
plosives	.87	.73	.62	.36
sonorants	.77	.64	.49	.17
total	.86	.75	.61	.36

Table 4: *Most likely confusion candidates for the ten most confusable onsets, figures indicate proportion of false hits.*

onset	Pinyin example	proportion correct	likely confusion candidates
w	**w**ang	.33	f (.21), l (.08), r (.08)
l	**l**ong	.34	r (.14), y (.08), t (.07)
∅	ang	36	r (.18), t (.11)
k’	**k**ong	.41	h (.16), k(.10), t (.09)
m	**m**eng	.47	p (.10), n (.10), r (.09)
n	**n**ong	.51	r (.18), y (.13), l (.10)
k	**g**ong	.56	h (.08), t (.07), y (.07)
h	**h**ong	.60	k (.06)
p	**b**ang	.63	h (.06), t (.06)
ts	**c**ong	.63	tS (.18), tC (.09)
t	**d**ong	.63	h (.06), r (.06)

Table 5: *Proportion of correct tone identification depending on the type of the nuclear vowel.*

nuclear vowel	proportion correct tone	Pinyin example
@	,61	sh**e**ng
a	,63	sh**a**n
e	,66	sh**e**i
E	,71	ni**a**n
i	,77	n**i**
I	,75	sh**i**
o	,54	p**o**
u	,66	sh**u**
U	,52	s**o**ng
y	,88	y**u**
total	,64	

We finally examined whether the type of the nuclear vowel influenced the tonal recognition. The result of this analysis is displayed in Table 5. Note that we only refer to the tone-bearing vowel part, even in diphthongs. As can be seen, front vowels such as [i] and [I] reach much higher recognition rates than back vowels. This suggests that the second formant frequency plays an important role in the process of tone recognition.

5. Discussion and Conclusions

In this study we examined the effect of added noise in the perception of mono-syllabic words of Mandarin. In a perception experiment participants were asked to type which Pinyin syllable and tone they had perceived.
It must be stated that the amount of data and the number of subjects is relatively small and therefore only allow tentative conclusions.

Our results suggest that the tonal information is more robust against noise than segmental information. Although each tone has only three counterparts to be confused with and the potential of segments to be confused is much higher, our results show that actual confusions only occur with very few candidates that in the case of consonants either share manner or place of articulation with a segment.

We found that the nuclear vowel is the part of the syllable with the highest recognition rate, followed by onset and coda.

In contrast to one of our earlier studies [3] the video image did not contribute to tone identification, but only enhanced segment recognition. This result may be explained by the different experiment design. In [3] participants were asked to identify the tone by choosing from a list of Chinese characters that all had the same segmental content. Therefore they did not experience the same cognitive load as in the current study where they had to identify both the tone AND the segments.

Furthermore we examined the influence of the syllable structure on tone perception and found that front vowels facilitate recognition whereas back vowel yield lower rates.

Future efforts will be dedicated to the perception of poly-syllabic words in a larger scale study.

ACKNOWLEDGMENTS

I would like to thank my Chinese students Xin Wang and Jiahui Fan who assisted me in recording the stimuli for this study.

REFERENCES

[1] Burnham, D., Ciocca, V., & Stokes, S., 2001. Auditory-visual perception of lexical tone, in Proceedings of Eurospeech 2001, Aalborg, Denmark, 395-398, 2001.
[2] Burnham, D., Lau, S., Tam, H., & Schoknecht, C. Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers, in Massaro, D., Light, J., & Geraci, K. (Eds) *Proceedings of AVSP2001*, pp 155-160, 2001.
[3] Mixdorff, H., Hu, Y. and Burnham, D., 2005. Visual Cues in Mandarin Tone Perception. In *Proceedings of Eurospeech 2005*, pp. 405 - 408, Lisbon, Portugal.
[4] http://www.virtualdub.org
[5] http://www.praat.org
[6] http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm
[7] http://pinyin.info