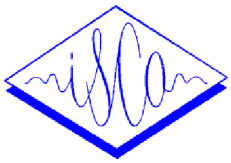


Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

AUDITORY-VISUAL PERCEPTION OF SYLLABIC TONES IN THAI

Hansjörg Mixdorff, Patavee Charnvivit** and Denis Burnham****

*Faculty of Computer Science and Media, TFH Berlin University of Applied Sciences, Germany

**Centre for Research in Speech and Language Processing, Chulalongkorn University, Bangkok, Thailand

***MARCS Auditory Laboratories, University of Western Sydney, Australia

ABSTRACT

This paper presents results concerning the use of visual cues in the perception of Thai tones. The lower part of a female speaker's face was recorded on digital video as she uttered 24 sets of syllabic tokens covering the five different tones of Thai. A perception study was conducted in which audio sound track alone; as well as audio plus video were presented to native Thai speakers who were required to decide which tone they perceived. Audio was presented in various conditions: clear, pink-noise masked at different SNR levels, and devoiced conditions using LPC resynthesis. Some subjects were presented with a video only, silent condition. In the devoiced and the clear audio conditions, there was little augmentation due to the addition of video. However, the addition of visual information significantly improved perception in the pink-noise masked condition, and the effect increased with decreasing SNR. Results on video only are close to chance suggesting that the improvement in noise-masked conditions is not due to additional information in the video *per se*, but rather to an effect of early integration of acoustic and visual cues facilitating auditory-visual speech perception.

1. INTRODUCTION

Syllabic tones in tone languages have distinct *F0* patterns. Thai has five different lexical tones: three static tones - mid (0), low (1) and high (3); and two dynamic tones, falling (2) and rising (4) (commonly used tone indices are given in brackets).

Research has shown that tone contours patterns in Thai and other tone languages can be associated with underlying tone commands of the Fujisaki model [1][2]. 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 [3] [4]. In addition, a preliminary study, by two of the current authors with digital video data was conducted to examine whether cues to the underlying tones might be present in the facial image [5]. Results included, *inter alia*, that tone identification in a video-only condition was difficult and depended on the syllables employed and the tones contrasted. Nevertheless, under certain cases identification was much better than chance. This observation plus the similar results found previously [8]] provided the impetus for the current study.

There has been so far relatively limited research with respect to the integration of acoustic and visual information in the perception of tone. An early EMG study [6] suggests that each tone of Thai is connected with distinct enervation patterns of the muscles involved. In the context of the current study, the behavior of extrinsic muscles controlling *F0* is of special interest, as these so-called strap muscles are directly connected with the articulatory system, that is, the muscles of the jaw and tongue. Physiological studies [7] also suggest certain restrictions with respect to the coordination of the laryngeal and articulatory systems which might be responsible for visual cues of tones.

In the associated realm of prosody, it has been shown that there is a strong correlation between head movements and F0 [8]. These correlations are continuous and seem to be used by multimodal perceivers during auditory-visual perception [9], but direct studies on the perception of these movements are yet to be conducted.

The current paper focuses on perception tests regarding Thai that were conducted in the context of a larger study of Thai, Vietnamese and Mandarin. For each of these languages mono-syllabic corpora were collected that consist of audio and video recordings as well as OPTOTRAK recordings with the associated audio (for later production studies). Only the perception data for Thai stimuli are presented here.

2. VIDEO MATERIAL AND STIMULUS MATERIALS

The corpus of monosyllabic tokens compiled for the current experiment contains a total of 24 syllables which were uttered by a female native speaker of Thai with the five different tones. The 24 syllables were chosen based on the following criteria: (1) a maximum number of members in each syllable set should represent real words, (2) a good coverage of Thai vowels, as well as articulatory trajectories (for instance, tongue movements from the back to the front, from the front to the back, etc.). All syllables are so-called "live" syllables with a sonorant coda. A list of the syllables used is shown in Table 1.

cun	laŋ	lu:aŋ	pɯŋ
ja:w	law	ma:j	sɔ:n
jiŋ	lim	man	seŋ
jo:ŋ	liw	muj	si:aw
k^hi:an	lom	mu:aj	wa:j
k^hlum	lon	ŋaw	wa:n

Table 1. List of syllables used in the study, IPA notation.

The tokens were randomized and recorded at MARCS Auditory Labs with a SONY video camera at DV standard (720 x 576 pixels, 25 frames per second) four times each. The audio was recorded using the standard Lavalier microphone shipped with the camera. The video sequences were then transferred onto a PC, together with the associated audio (48 kHz, 16 bit).

In order to segment the long video sequences into chunks of individual tokens, the audio tracks were downsampled to 16 kHz and annotated using *Praat TextGrid* [10]. A tool was written for converting the *TextGrid* into a *VirtualDub* [11] script which in turn was 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 onset of the syllable and ending 400 ms after the offset.

In order to determine the potential contribution of visual cues to tone perception two different audio degradation paradigms were adopted:

1. Reduced devoiced audio, that is, stimuli which presumably do not contain acoustic tonal information

2. Masked audio, that is, the original audio masked by varying levels of noise

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

Accordingly, the original soundtracks were subjected to the following manipulations.

Devoiced stimuli were created by LPC analysis in *Praat* (default settings: Prediction order 16, window-size 25 ms, step-size 5 ms) and resynthesis using pink noise as the source signal. Pink noise was chosen rather than white noise, as the resulting speech stimuli were more similar to whispered speech and more comfortable to listen to.

Pink-noise masked stimuli were created by adding pink noise with a decaying spectral slope of 6 dB/octave at SNRs of -15, -16.5, 18, 19.5 and -21 dB, the SNR being calculated for the speech portion only, not the silent intervals before and after. The appropriate values of SNR were determined by preliminary trials, identifying the region between tone identification performances on clear-audio and complete masking.

A *VirtualDub* script was used for replacing the original soundtracks by clear 16 kHz, pink-noise masked and devoiced versions. From each syllable, two example tokens were selected, yielding 24 syllables x 5 tones x 2 tokens = 240 stimuli of each version. For easier reference we introduce abbreviations for the types of stimuli: clean 16kHz audio, **Clean-A**, clean 16 kHz audio plus video, **Clean-AV**, devoiced audio, **DeVoiced-A**, devoiced audio plus video, **DeVoiced-AV,** pink-noise masked audio, SNR=-15dB, **Noise15-A,** pink-noise masked

audio plus video, SNR=-15dB, **Noise15-AV**, etc. Video only stimuli are indicated by **VO**.

3. THE PERCEPTION TEST

Experiments were conducted using the *DMDX* software [12] and employed scripts provided by Caroline Jones (MARCS, UWS) that were slightly modified. Considering the large number of stimuli and the fact that the tests were to be conducted with native Thai speakers, an identification task rather than a discrimination task was employed. This required participants to identify the presented stimulus by choosing one of five written syllables differing only in tone.

One set of syllables was chosen for a practice session preceding the experiment proper, and the remaining 23 syllable sets were divided into four groups. During a session each subject was presented with stimuli from four different auditory, visual 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, **Clean-A**, **Clean-AV**, **Devoiced-A**, and **Devoiced-AV,** with each block containing a different set of syllables, and the sequence of stimulus types varying between the four trial sets.

Block	Trial set 1	Trial set 2	Trial set 3	Trial set 4
1	Stim.type 1 syll.group 1	Stim.type 2 syll.group 1	Stim.type 3 syll.group 1	Stim.type 4 syll.group 1
2	Stim.type 2 syll.group 4	Stim.type 1 syll.group 2	Stim.type 2 syll.group 2	Stim.type 3 syll.group 2
3	Stim.type 3 syll.group 4	Stim.type 4 syll.group 3	Stim.type 4 syll group 4	Stim.type 1 syll.group 4
4	Stim.type 4 syll.group 2	Stim.type 3 syll.group 4	Stim.type 1 syll.group 3	Stim.type 2 syll.group 3

Table 2: Structure of experiment trials with respect to stimulus type and syllable set.

Within each block, tokens pertaining to a syllable set were presented consecutively, but in randomized order. Each combination of syllable and tone occurred four times, that is, each of two different examples of the syllable/tone combination was presented twice. Hence each trial set consisted of 23 syllables x 5 tones x 2 versions x 2 repetitions=460 trials and took about 45 minutes to complete.

Participants listened to the stimuli over headphones connected to a PC soundcard. Each trial started with a preparation phase of one second during which the word 'ready' was displayed. Then the stimulus was presented, followed by the five Thai syllables written in Thai script. The order of the five syllables corresponded to the numbering conventions for Thai tones and was left unaltered during the experiment. Following the presentation of these five syllables, participants made a forced choice by hitting the appropriate number key on the keyboard.

In the practice trials (using one syllable set), feedback concerning response accuracy was given, but in the main test no feedback was given.

Participants were eight staff of Centre for Research in Speech and Language Processing (CRSLP), Chulalongkorn University, Bangkok, aged 22-34, and a total of 40 students of the same university aged between 18 and 23. They reported normal hearing, and two had corrected vision. None of them were familiar with the speaker who had produced the video data. The number of participants varied between the trial sets.

4. RESULTS

The results are discussed in two parts, first with respect to conditions employing clean and devoiced audio stimuli, and second with respect to the masked audio and silent video.

4.1 Clean and Devoiced Audio Stimuli

In the clean and devoiced audio stimulus conditions, eight participants (3 males, 5 females), all staff at CRSLP, took part. **Figure 2** displays the proportion of correct responses for stimulus types **Clean-A**, **Clean-AV**, **DeVoiced-A**, and **DeVoiced-AV**.

Given that the number of possible tones is five, the statistical chance level for the current experiment is at 20%. As can be seen, the identification rate on clear audio is close to 100%. The average results suggest only slight differences between audio only and the corresponding audio plus video conditions. Marginal gains (especially for the low tone, number 1) can be observed in the **Clean-AV** condition ($p<.05$). In the case of the devoiced stimuli the picture is rather mixed. Whereas the falling tone (2) is recognized only around chance level, the rising tone (4) yields up to 71% correct responses. The corresponding tonal confusion matrix for the **DeVoiced-AV** condition is displayed in Table 3.

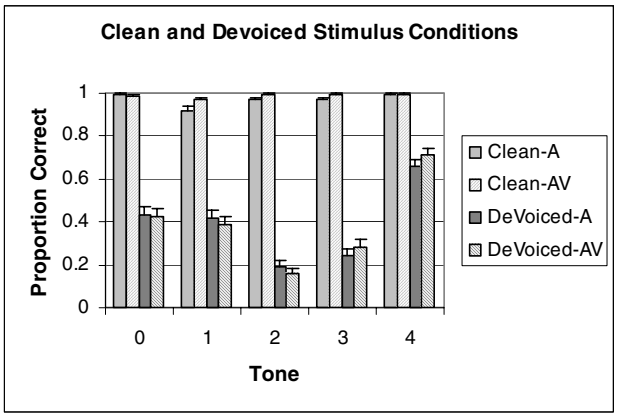


Figure 2. Results of perception experiment employing stimuli types Clean-A, Clean-AV, DeVoiced-A, and DeVoiced-AV.

	Perceived tone				
Intended tone	0	1	2	3	4
0	**42.39**	27.72	5.43	16.85	7.61
1	33.70	**38.59**	12.50	5.98	9.24
2	41.85	29.89	**15.76**	7.07	5.43
3	24.46	11.96	8.15	**28.26**	27.17
4	10.33	8.15	1.09	9.24	**71.20**

Table 3. Confusion matrix for stimulus type **DeVoiced-AV**. Rows indicate intended tone, columns perceived tone, figures correspond to percent.

As can be seen, tones 0 and 1 are highly confusable as well as tones 3 and 4, though the confusion in the latter two primarily occurs in one direction. We can translate this matrix into a matrix of perceptual distance by calculating the mean probability that confusion between two tones *i* and *j* does not occur, given by:

$$p_{correct} = 1.0 - (p(i \rightarrow j)+p(j \rightarrow i))/2.$$

As can be seen from the matrix in Table 3 confusions are not necessarily symmetrical. $p_{correct}$ for condition **DeVoiced-AV** is displayed in Table 4.

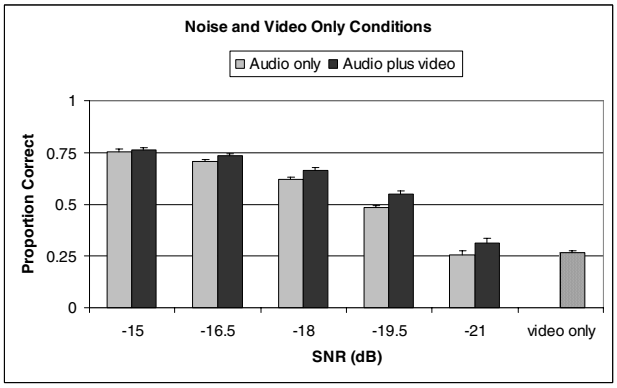
	Tone *j*			
Tone *i*	0	1	2	3
1	69.2			
2	76.4	78.7		
3	78.8	91.0	92.4	
4	91.0	91.3	96.7	81.8

Table 4. Perceptual distances $p_{correct}$ between the five syllabic tones, condition **DeVoiced-AV**.

4.2 Pink Noise Masked Audio Stimuli

The experiment series using pink-noise masked stimuli and silent video involved a total of 40 students of Chulalongkorn university (24 males and 16 females). 16 subjects performed in the tests at SNR levels between -15 and -19.5 dB, and four in the test at -21 dB and the silent videos.

Figure 3 (top) displays the proportion correct for pink-noise masked stimuli types**,** as well as video only. Due to space limitations results for individual tones are only displayed for **Noise19.5-A** and **Noise19.5-AV** as shown in Figure 3 (bottom). As can be seen from the top figure, recognition rates are still fairly high at an SNR of -15 dB, and the subjects do not significantly benefit from seeing the video. As the SNR decreases, however, the relative gain of the "plus video" conditions increases, from 4.1% at -16.5dB to 14.0% at 19.5 dB. The gain is almost significant at -16.5dB (p=.052), significant at -18dB (p=.03), and highly significant (p<.01) for -19.5dB. The result for **VO** is slightly above chance level.



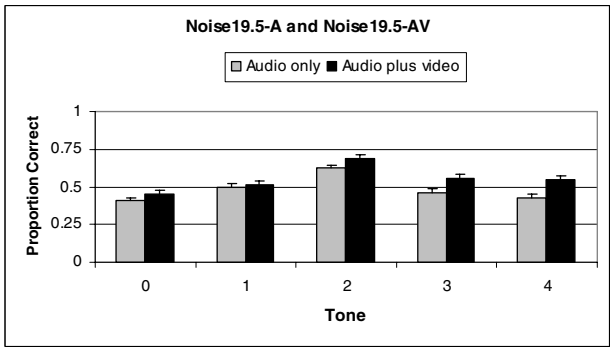


Figure 3. Top: Pooled results of perception experiments employing stimuli types Noise15-A, Noise15-AV, Noise16.5-A, Noise16.5-AV, Noise18-A, Noise18-AV, Noise19.5-A, Noise19.5-AV, Noise21-A, Noise21-AV and VO. Bottom: Results for Noise19.5-A and Noise19.5-AV depending on the tone.

The recognition results for individual tones are much more balanced than in the **DeVoiced-AV** condition, as is confirmed by the confusion matrix for case **Noise19.5-AV** shown in Table 5.

	Perceived tone				
Intended tone	0	1	2	3	4
0	**45.15**	14.54	5.10	20.41	14.80
1	10.71	**51.02**	6.63	9.44	22.19
2	3.57	7.14	**69.13**	15.05	5.10
3	19.39	7.65	6.12	**55.36**	11.48
4	4.08	25.51	4.34	10.97	**55.10**

Table 5. Confusion matrix for stimulus type **Noise19.5-AV**. Rows indicate intended tone, columns perceived tone, figures correspond to percent.

The corresponding $p_{correct}$ are displayed in Table 6. They suggest that tones 0 and 3, as well as tones 1 and 4 are most prone to confusion. This is interesting in the light of auditory-only studies of confusion of Thai tones. For example, Abramson [13] found that in Thai the mid (0) and low (1) tones were most confusable in an auditory-only test. It may be the case then that some confusions are resolved to some extent by the addition of auditory-visual cues.

	Tone *j*			
Tone *i*	0	1	2	3
1	87.4			
2	95.3	93.1		
3	81.1	91.5	89.4	
4	90.6	76.1	95.3	88.8

Table 6. Perceptual distances $p_{correct}$ between the five syllabic tones, condition **Noise19.5-AV**.

The results of the masking experiment confirm earlier observations [4] that presenting the video along with the masked audio increases the tone identification rate. This gain appears to increase with increased masking of the speech sound. At an SNR of -19.5dB most syllables themselves become very difficult to identify while the tonal contour can still be detected, as informal listening demonstrated. This, is, of course, the fundamental difference between the pink noise stimuli and the devoiced stimuli which are devoid of tonal information and do not significantly benefit from the addition of video. It is possible that the video does not contain any additional information in and of itself that the subjects can use. This, however, does not necessarily imply, that the tones do not have specific facial correlates. These could well exist, but are either not captured by the video or cannot be exploited by the perceiver. The cues for tone may reside not simply in the video alone but in the articulatory gesture, an auditory-visual event. The extent to which the time-varying cues in the auditory and visual information line up, may be the extent to which there is auditory-visual augmentation over auditory alone speech perception [14][15][16]. The auditory noise stimuli have such information, and the AV conditions show augmentation above the audio alone. On the other hand, identification of tones on the devoiced stimuli was much poorer and any auditory-visual augmentation may be due to the presence of some fundamental frequency and intensity cues over time. Tone 4 in Thai, for instance, exhibits a characteristic two-peaked intensity contour as compared to other tones (see Figure 4).

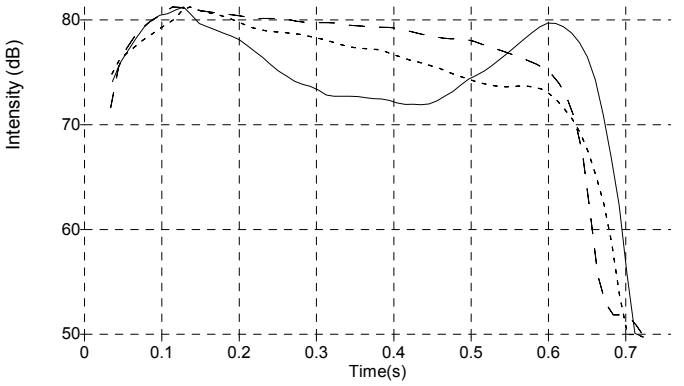


Figure 4. Typical intensity contours of the syllable [maaj] for mid tone 0 (dashed), low tone 1 (dotted) and rising tone 4 calculated and drawn with *Praat*.

5. DISCUSSION AND CONCLUSIONS

The perception experiment shows that under clear audio conditions correct identification rates were close to 100% and improved only slightly when the video was presented at the same time, an obvious ceiling effect.

Identification rate on devoiced audio was reduced to a mean of 40%, but did not significantly improve with the addition of video. Nevertheless, there were some above-chance levels of identification, which may be explained by specific temporal

characteristics (duration, intensity) which were still present in the devoiced stimuli for some of the tones. Furthermore, a recent vowel study [17] showed a correlation between f0 and F1 which may suggest that a certain amount of tonal information may be captured in the spectrum envelope.

There were significant benefits from presenting the video with pink-noise masked stimuli, and this gain increased with decreasing SNR. Thus it can be concluded that there is augmentation of tone perception in auditory-visual compared to auditory alone conditions, especially when the tone contour is retained in the auditory signal.

Ultimately, articulatory gestures and their facial correlates are primarily a function of the syllable being uttered and only secondarily a function of tone. The fact that there was auditory-visual augmentation in the noise-masked conditions suggests that the benefits of the video are due to an effect of early auditory-visual integration of the time varying and modality-independent characteristics of the tone even when the underlying syllable may no longer be identified. This, however, does not imply that visual-only cues for tones do not exist; they may not be captured by the video due to limited temporal and spatial resolution. Normal hearing perceivers might simply not be trained to make use of tone cues, since in a typical communicative situation; success hardly depends only upon correctly perceiving individual tones, as the discourse is embedded in a rich context of topic and non-verbal gestures. Studies like the one presented here nevertheless illustrate how auditory-visual cues operate at a basic level to augment speech perception.

Future efforts will be dedicated to a more detailed analysis of perception test results with respect to the syllable types, evaluation of acoustic features of the corpus data, evaluation of OPTOTRAK data to determine the articulatory concomitants of tone, and cross-language comparisons with the other language corpora that have been collected.

6. ACKNOWLEDGMENTS

This work was supported by a grant from the University of Western Sydney International Research Initiatives Scheme, as well as a DAAD short-term lectureship, grant D/04/01405. The assistance of Dr Caroline Jones in providing DMDX scripts and of Rua Haszard Morris and Colin Schoknecht in running experiments is gratefully appreciated.

7. REFERENCES

[1]. Fujisaki, H.; Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences *of Japanese"*. *Journal of the Acoustical Society of Japan (E), 5(4),* 233-241, 1984.

[2]. Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. and P. Charnvivit, P. "Perception of Tone and Vowel Quantity in Thai. *Proceedings of ICSLP 2002*, Denver, USA, 2002

[3]. Burnham, D., Ciocca, V., & Stokes, S., 2001. "Auditory-visual perception of lexical tone," in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 395-398, 2001.

[4]. 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 Auditory-Visual Speech Perception Conference 2001 (AVSP2001)*, Causal Productions, www.causal.on.net, 155-160, 2001.

[5]. Mixdorff, H. and Charnvivit, P. "Visual Cues in Thai Tone recognition," Proceedings of TAL 2004, pp. 143-146, Beijing, China, 2004.

[6]. Erickson, D., *A Physiological Analysis of the Tones of Thai*. PhD thesis, University of Connecticut, 1976.

[7]. Xu, Y. and Sun, X. "Maximum speed of pitch change and how it may relate to speech," *Journal of the Acoustical Society of America* 111: 1399-1413, 2002.

[8]. Yehia, H.C., Kuratate, T., & Vatikiotis-Bateson, E. "Linking facial animation, head motion and speech acoustics," Journal *of Phonetics*, *30*, 555-568, 2002.

[9]. Vatiktiotis-Bateson, E., Kroos, C., Kuratate, T., Munhall, K. G., & Pitermann, M. Task constraints on robot realism: The case of talking heads. In K. Kamejima (Ed.), *9th IEEE International Workshop on Robot & Human Interactive Communication (RO-MAN 2000),* (pp. 352-357). Osaka, Japan: IEEE, 2000.

[10]. http://www.praat.org

[11]. http://www.virtualdub.org

[12]. http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm

[13]. Abramson, A. S. *The Thai tonal space*. *Haskins Laboratories: Status report on Speech Research*, 85, 105-114, 1986.

[14]. Davis, C.& Kim, J. "Audio-visual interactions with intact clearly audible speech," *Quarterly Journal of Experimental Psychology*, 2004.

[15]. Kim, J. & Davis, C. "Hearing foreign voices: does knowing what is said affect masked visual speech detection? *Perception*, 32, 111-120, 2003.

[16]. Kim, J., Davis, C., Vignali, G., & Hill, H. Visual speech concomitants of the Lomard reflex. *Proceedings of AVSP 2005*, Causal, 2005, In this volume.

[17]. Pfitzinger, H.R."Acoustic Correlates of the IPA Vowel Diagram," *Proceedings. ICPhS 2003*, vol. 2, pp. 1441-1444. Barcelona, 2003.