The Perception of Norwegian Word Tones by Chinese and German Listeners

Wim A. van Dommelen, Olaf Husby

Norwegian University of Science and Technology, Norway
wim.van.dommelen@hf.ntnu.no, olaf.husby@hf.ntnu.no

1. Introduction

In contrast to almost all other European languages, Norwegian to a certain extent exploits lexical tones in its sound system. The phonological system involves two different tones (tone 1 and tone 2) that are used to distinguish a relatively large number of word pairs which are only distinguished by their lexical tone. An example is the word pair *loven* (/^{1}loːvən/ ['the law']) and *låven* (/^{2}loːvən/ ['the barn']). The Norwegian word tones can be realized in different ways, depending on the speaker's dialect (Fintoft, 1987; Kristoffersen, 2000:233-238). In the present study, we shall deal with a so-called low-tone dialect having tones that are phonologically described as LH (tone 1) and HLH (tone 2; see the example given in Figure 1).

It is obvious that for users of Norwegian as a second language (L2) the lexical tone system poses an extra challenge beyond the acquisition of the segmental aspects. From a theoretical point of view, it is difficult to make predictions about L2 behavior regarding word tones. Current models of the acquisition of the phonetics of an L2 have been developed for the segmental rather than the suprasegmental domain (Perceptual Assimilation Model; Best, 1995; Speech Learning Model; Flege, 1995). Most studies on specific issues related to L2 acquisition deal with perception at the phoneme level (e.g., Kingston, 2003; Pruitt, Jenkins & Strange, 2006) or the word/sentence level (e.g., Rogers, Dalby & Nishi, 2004; Van Engen & Bradlow, 2007). Though L2 tone issues have been investigated less, previous work has already shown that L2 users of tone languages have difficulties in both perceiving and producing tones correctly (for an overview on Mandarin tone acquisition, see Wang, Jongman & Sereno, 2006). In addition, native speakers of a tone language have been shown to encounter difficulties in interpreting prosodic contours of a nontonal language (Yudong, 2007).

The goal of the present study was to investigate the perception of the Norwegian tones by L2 listeners. To that aim, speakers from two different language types, namely a tone language (Mandarin Chinese) and a nontonal language (German) were included. It was hypothesized that speakers from the former language type would generally perform better than those from the latter (cf. Kaan, Wayland, Bao & Barkley, 2007). Subjects were recruited who had only recently come to Norway and started learning the language in a university course. After having attended a general lecture on Norwegian pronunciation issues, all subjects participated in a listening test involving an AXB discrimination task and a test on identification of the two Norwegian tones (for details, see Section 2 and Table 2). After this test session, both the Chinese and the German listener group were split up into two subgroups. One subgroup attended a training session on the Norwegian tones as used in the listening test, while the other subgroup was offered a general lecture on phonetic issues in Norwegian excluding lexical tones. Following these lectures all the subjects performed a discrimination test similar to the first one and a repetition of the identification test. It was expected that the listeners who received tonal training would show improved discrimination as well as identification performance. Further, it was speculated that the Chinese listeners would profit from their experience with tones in their native language and show a larger improvement than the Germans.

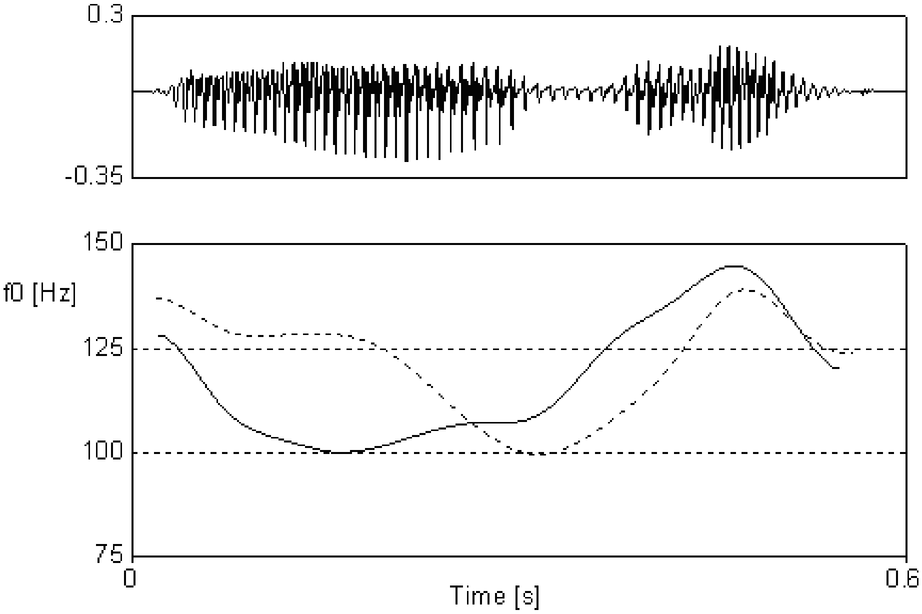

Figure 1. Waveform of stimulus word /loːvən/ with original tonal contour LH (tone 1; solid line) and substitute contour HLH (tone 2; dashed line).

2. Experimental procedure

2.1 Speech material

The speech material used for this study consisted of two parts. Part one was a one-page text especially written to have the character of a novel fragment. Contained in the text were nine different tone 1 words and their tone 2 counterparts, thus yielding a total of 18 stimulus words. The stimulus words occurred in a natural way without any systematicity. The second part of the speech material consisted of isolated words that were elicited in a dialogue and spoken with a focal accent. Here, the speakers were asked to respond to questions like “What is the definite singular form of [stimulus word]?” The same nine pairs of tone 1 and 2 words were recorded, i.e., a total of 18 isolated words.

Using this speech material, two low-tone dialect speakers were recorded: a 25-year-old female speaker and a male speaker aged 26. At the time of the recordings, both speakers were students at NTNU's Department of Scandinavian Studies and Comparative Literature. The recordings were made in a sound-treated studio using a Milab LSR 1000 microphone and a Fostex D-10 digital recorder and were stored on hard-disk with a sampling rate of 44.1 kHz. Speech signal editing and manipulations were carried out using Praat (Boersma & Weenink, 2007).

2.2 Discrimination test

The following procedure was used to generate the stimulus material for the discrimination test. In each of the 2 (speakers) x 9 (words) x 2 (tones)= 36 tokens spoken in isolation the original fundamental frequency contour was replaced by its counterpart, thus substituting tone 1 by tone 2 and *vice versa*. Substitutions were achieved with the help of Praat's *Replace pitch tier* facility with some small corrections in both the time and the fundamental frequency domain that were necessary due to differences between tone 1 and tone 2 originals. In this way, manipulated versions were created that differed only in f_0 contour but not in other parameters (as illustrated in Figure 1). According to informal listening results, it was not possible to tell manipulated from original versions in tone 1/tone 2 word pairs.

AXB stimuli were prepared by combining three tokens each into an audio file with pauses of 750 ms in between. Each original token resulted in four different AXB stimuli. Taking an original tone 1 token combined with its manipulated tone 2 counterpart as an example, the following AXB stimuli were created: 1-1-2, 2-1-1, 2-2-1, and 1-2-2 (see Table 1). Thus, a total of 36 (originals) x 4 (combinations)= 144 different AXB stimuli were generated. The stimuli were presented to the listeners in discrimination tests *Pre* and *Post* containing 72 stimuli each (see also 2.5 Listening test design). For each listener, a different randomized order was generated.

Table 1. Format of AXB stimuli in discrimination tests *Pre* and *Post*. 1 (orig)= test word with original tone 1 contour; 2 (man)= manipulated test word with tone 1 contour substituted by tone 2 contour. Similarly for original contour 2 replaced by contour 1.

Discrimination test	**A**	**X**	**B**
Pre	1 (orig)	1 (orig)	2 (man)
	2 (man)	1 (orig)	1 (orig)
	2 (orig)	2 (orig)	1 (man)
	1 (man)	2 (orig)	2 (orig)
Post	1 (man)	1 (man)	2 (orig)
	2 (orig)	1 (man)	1 (man)
	2 (man)	2 (man)	1 (orig)
	1 (orig)	2 (man)	2 (man)

2.3 Identification test

The stimulus material in the identification test comprised 72 different tokens in total: 2 (speakers) x 9 (words) x 2 (tones)= 36 tokens spoken in isolation and 36 similar tokens excised from the short story. For each listener, a different randomized order was generated. The identification was run twice (see 2.5 Listening test design).

2.4 Subjects

Two groups of subjects participated in the listening tests: 10 listeners with Mandarin Chinese and 11 with German as their native language. The age range of the former group was 22-30 years (mean 26.7) and of the latter 21-28 years (mean 23.0). None of the subjects suffered from hearing impairments. All listeners were recruited from the same basic course in Norwegian offered by the Norwegian University of Science and Technology. Whereas the duration of their stay in Norway did not really vary among the members of the German group (ranging from three to four months), the group of Chinese was less homogeneous in that respect (mean duration of 11 months; ranges from 3-6 months for four subjects, 10-16 months for five subjects and a duration of 26 months for one subject). The listeners were paid for their participation.

2.5 Listening test design

Preceding the first listening test session, all subjects participated in a 45-minutes lecture on Norwegian pronunciation (see Table 2). Topics dealt with in this lecture were, among other things, the Norwegian vowel and consonant system, syllable structure, and stress. Also the tone system was mentioned and briefly demonstrated. Shortly after the lecture (with delays varying between half a day and a few days) discrimination test *Pre* and the identification test (see above) were run. Individual listeners were presented with the stimuli over high-quality loudspeakers and responded by mouse-clicking response alternatives presented on a computer screen. The listener's response prompted the next stimulus to be presented. In the AXB test, the listeners' task was to indicate whether X was judged to be identical with A or B by clicking one of the response alternatives *B= A* and *B= C*. Response alternatives in identification were *tone 1* and *tone 2*. Written instructions were given for each of the two tests. In addition, a few tone 1/tone 2 word pairs were demonstrated.

Table 2. Listening test design. Stimuli in discrimination tests *Pre* and *Post* were different, but had the same design (see text). Stimuli in the identification test were identical *Pre-Post*.

	Listener group	
	Training	**No Training**
Lecture on Norwegian pronunciation	X	X
Discrimination test *Pre*; identification test	X	X
Lecture on Norwegian tones	X	
General lecture		X
Discrimination test *Post*; identification test	X	X

One week after the first lecture, two further 45-minutes lectures were held. One of them dealt with the realization of Norwegian tones. Tone 1/tone 2 word pairs were demonstrated by a phonetically trained female speaker of a low-tone dialect and were imitated by the participants under her guidance. Five of the Chinese and six of the German listeners attended this lecture. Henceforth, these 11 subjects will be called the *Training* group. The ten remaining subjects attended the other lecture, which was of a general character and did not include any mention of the tone system. These subjects formed the so-called *No Training* group. Similar to the procedure followed for the first test session, the subjects took part in discrimination test *Post* and a repetition of the identification test.

2.6 Evaluation

Evaluation of the discrimination test data involved registration of the listeners' responses as correct or incorrect. The results will be expressed in percent correct. Chance level is 50 % correct. The data from the identification tests could not be evaluated in their raw form. The reason was that some listeners obtained less than 50 % correct identification. Since the task of the listeners was to attach one of the (subjectively rather) arbitrary labels *tone 1* or *tone 2* to a stimulus, a raw identification rate of e.g., 35 % correct can be interpreted as a 65 % correct recognition of a pattern. Thus, it was decided to recompute all cases of mean identification rates of less than 50 % by the formula x= 100 – raw %.

3. Results

3.1 Discrimination of word tones

The results of discrimination tests *Pre* and *Post* are presented in Figure 2. First of all it can be noted that in general discrimination performance was high, clearly above chance level (50 %). Not unexpectedly, overall Chinese performance was better than that for the Germans, however, not dramatically so (97.3 % *vs.* 85.8 %; statistically significant, p<0.001; see Table 3). Further, the data showed a general increase in performance in the *Post*-test *vs.* the *Pre*-test. Pooled across the listener groups, the already high discrimination rate of 88.9 % correct in the *Pre* condition was still significantly higher for the *Post* session (93.7 %; p<0.001). Closer inspection of the data revealed that the *Pre*/*Post* effect is mainly due to increased discrimination rates for the Germans. Whereas the increase for the latter group pooled across the two training conditions amounted to 7.7 % (according to a t-test for independent samples highly significant; t(1582)= -4.42; p< 0.001), the corresponding amount for the former group was only 1.5 % (statistically non-significant). This lack of improvement is simply a ceiling effect (performance rates of 96.5 % correct for *Pre* and 98.0 % for *Post*).

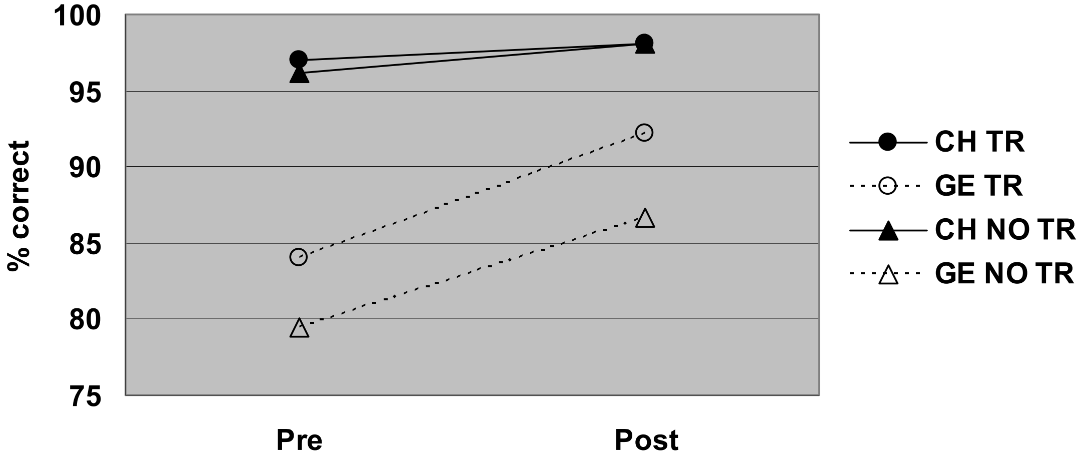

Figure 2. Percent correct discrimination rates for tests *Pre* and *Post* performed by Chinese (CH) and German (GE) listener groups with (TR) and without training (NO TR).

Splitting up the *Pre*/*Post* improvement of 7.7 % for the German listeners according to the factor training showed similar values for the *Training* and the *No Training* group (92.1 - 84.0= 8.1 % and 86.7-79.4= 7.3 %, respectively; see Figure 2). According to a separate ANOVA for the German subjects there was no significant interaction between the factors *Pre/Post* and *Training* (F(1, 1576)= 0.06; p=0.801). The generally better performance of both the *Training* and the *No Training* group in the *Post* condition turned out to be significant (F(1, 1576)= 8.25; p= 0.004).

The increased performance in spite of the lack of training on tone for the latter group suggests that for both groups the improvement might be due to familiarization with the test format rather than phonetic training. To investigate this, the response data from each of the listeners were divided into two halves. Improved performance in the second half of the discrimination test (in particular, in the *Pre*-test) would indicate the presence of a familiarization effect. The data revealed generally rather small differences between discrimination rates for the second *vs.* the first half, ranging from -0.3 % for the Chinese listeners in the *Pre* condition to 1.3 % for the German subjects in the same condition. In the *Post* condition the differences were smaller still. The ANOVA presented in Table 3 showed that the familiarization factor indeed did not have any significant effect (p= 0.439). The same was true for the interactions L1 x Familiarization and L1 x Pre/Post x Familiarization (p= 0.525 and p= 0.755, respectively).

Table 3. Analysis of variance for discrimination tests *Pre* and *Post*. Effect of the factors L1 (*Chinese*, *German*), Training (*Training, No Training*), Pre/Post and Familiarization (see text). Not all interactions are reported.

Source of variation	F	df	p
L1	136.82	1, 3008	<0.001
Training	7.36	1, 3008	0.007
Pre/Post	21.01	1, 3008	<0.001
Familiarization	0.60	1, 3008	0.439
L1 x Training	5.28	1, 3008	0.022
L1 x Pre/Post	9.36	1, 3008	0.002
L1 x Familiarization	0.40	1, 3008	0.525
L1 x Pre/Post x Familiarization	0.10	1, 3008	0.755

3.2 Identification of word tones

From Figure 3 it can be seen that the scores for tone identification were substantially lower than for discrimination. According to chi-square tests, the two lowest rates of 53.1 % and 53.6 % (for the German *No Training* group in the *Post*-test and the Chinese *Training* group in the *Pre*-test, respectively) were at chance level. The other scores were significantly above chance. The generally low performance was observed for Chinese and German listeners alike, with overall scores of 60.2 % correct and 56.1 % correct, respectively. According to an ANOVA with the factors L1, Training and Pre/Post this difference between the two groups reached statistical significance (p= 0.018; see Table 4). As to the effects of the conditions *Pre vs. Post* and *Training vs. No Training* the picture is more complicated than for the discrimination results. Let us start with performance of the Chinese listeners. Among them, the *No Training* group achieved a nearly 10 % higher mean score in the *Pre* condition than the *Training* group (63.3 % *vs.* 53.6 %; t(718)= 2.66; p= 0.008). In the *Post* condition, the performance of the former group remained virtually unchanged (62.8 %; non-significant). In contrast, the Chinese listener group that received training improved their performance significantly by 7.5 % from 53.6 % to 61.1 % (t(718)= -2.04; p= 0.042).

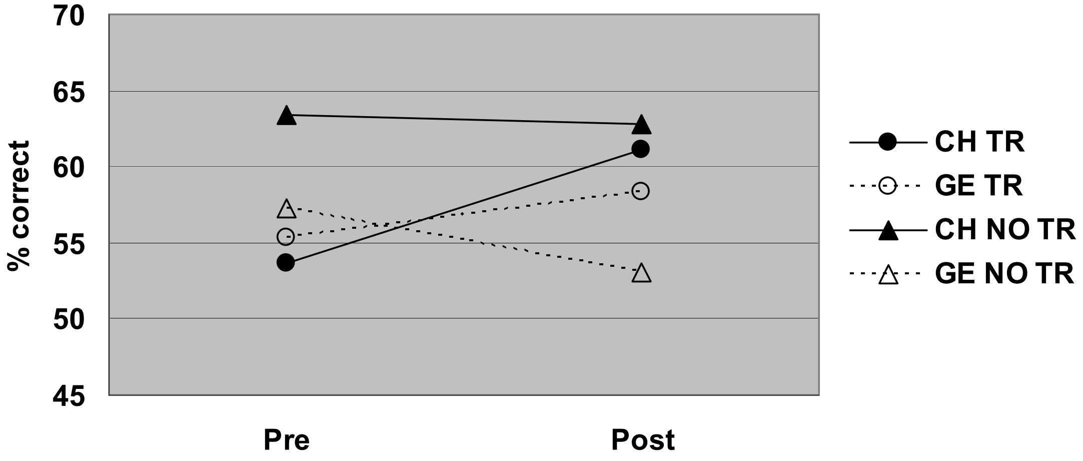

Figure 3. Percent correct identification rates for tests *Pre* and *Post* performed by Chinese (CH) and German (GE) listener groups with (TR) and without training (NO TR).

Table 4. Analysis of variance for identification tests *Pre* and *Post*. Effect of the factors L1 (*Chinese*, *German*), Training (*Training, No Training*) and Pre/Post. Not all interactions are reported.

Source of variation	F	df	p
L1	5.57	1, 2992	0.018
Training	1.25	1, 2992	0.264
Pre/Post	0.65	1, 2992	0.419
L1 x Training	4.25	1, 2992	0.039
L1 x Pre/Post	1.28	1, 2992	0.258
L1 x Training x Pre/Post	0.02	1, 2992	0.902

Different from the two Chinese subgroups, the German *Training* and *No Training* groups had similar identification rates in the *Pre* test condition (55.3 % and 57.2 %, respectively; a non-significant difference; t(790)= 0.54; p= 0.592). In the *Post* condition, the mean identification rate for the *No Training* group was somewhat reduced to 53.1 %. A t-test showed, however, that this reduction was non-significant (t(728)= 1.12; p= 0.262). Therefore, the *No Training* group's *Pre* and *Post* performance rates have to be regarded as the same. Finally, the German *Training* group achieved somewhat higher mean identification rates in the *Post* condition, the increase being from 55.3 % to 58.3 % correct. This 3 % increase appeared to be statistically non-significant (t(862)= -0.89; p= 0.372).

4. Discussion

The results of both the tone discrimination and identification tests partly confirmed and partly rebutted our experimental hypotheses. In line with our expectations regarding discrimination, the Chinese listeners outperformed the Germans with generally high percent correct rates. Discrimination performance for the latter group was, however, also substantially higher than chance level. These results might be due to the fact that the Norwegian tone system has merely two different tonal melodies. On the other hand, the tonal differences are relatively subtle so that listeners of a nontonal language might be expected to encounter difficulties in perception. A feasible explanation for the generally high German performance could be that these listeners judged the stimuli using psychoacoustic rather than linguistic criteria. This explanation is in line with the findings by Huang (2004:39) for AX discrimination of four tones in Chinese Putonghua. In that experiment, native speakers achieved on average between 91 % and 98 % correct and the rates for English listeners were similar, ranging between 85.5 % and 99 %. In an investigation by Hallé, Chang & Best (2004), Mandarin Chinese and French listeners performed discrimination tests on Mandarin Chinese tone continua. Mean discrimination rates for these two groups were 88 % and 74 %, respectively. Different from the Chinese subjects, the French listeners did not show enhanced discrimination sensitivity near category boundaries. Therefore, the authors argue that the perceptual behavior of the nontonal listeners seems to be determined by psychophysical factors. Our own results fully harmonize with the above-mentioned evidence and reasoning.

The present results showed improved discrimination performance for the German listeners in the *Post*-test condition. The absence of an improvement for the Chinese subjects could readily be explained as a ceiling effect. Unexpectedly, the behavior of the nontonal listeners is harder to interpret. On the one hand, the improvement for the German group that received tone perception training is in congruence with similar findings in previous studies (e.g., Wang, Spence, Jongman & Sereno, 1999). It would seem natural to explain this improvement as due to the training. But on the other hand, also the *No Training* group performed better in the *Post*-test, the size of the improvement being non-significantly different from the value measured for the *Training* group. Obviously, there does not have to be a causal relationship between perception training and improved performance. It was speculated that the alleged training effect could actually be familiarization with the listening test situation. Our analysis of the listeners' performance revealed, however, no difference between the first and second halves of each experimental session. Therefore, the familiarization hypothesis seems to be ruled out. What remains then is the possibility that our subjects had generally become more aware of the tonal aspects of Norwegian and so achieved an increased sensitivity.

In the identification task, listener scores were substantially lower than in discrimination. While one out of four mean scores in each of the test conditions *Pre* and *Post* appeared to be non-significantly different from chance level (50 %), the highest mean identification score amounted to 63.3 % (Chinese *No Training* group in the *Pre*-test). The present scores are clearly lower than, e.g., those found by Wang, Spence, Jongman & Sereno (1999) for the identification of four Mandarin tones by native speakers of American English. In the pre-test condition of that study, two groups of listeners achieved rates of 69 % correct (training group) and 67 % correct (control group).

The effect of training on identification performance turned out to be rather small. For the German group, scores in the *Post*-test were higher than in the *Pre*-test, but the size of the increase was only 3 % and statistically not significant. The improvement of 7.5 % found for the Chinese listeners, however, was robust. Though the effect of training thus was moderate, the general picture of better performance of the tonal *vs.* the nontonal listeners confirms our initial expectations. It seems reasonable to assume that the language-specific results are due to different perceptual mechanisms. Since the perceptual system of German listeners is lacking tonal categories, they can be considered 'tone-deaf'. Obviously, the training they received was not sufficient to remedy this situation. On the one hand, since the training comprised merely one lecture the lack of substantial improvement is understandable (cf. the mean improvement of 21 % for the identification of Mandarin tones after a two-week training program, consisting of eight sessions of 40 min each, in Wang, Spence, Jongman & Sereno, 1999). On the other hand, directly before actually performing the identification test some subjects gesturally demonstrated the typical movements of tone 1 and tone 2. Their (correct) demonstrations showed that they were aware of the relevant tonal properties. Nevertheless, inspection of individual data revealed that their results were not better than average. The rather small improvement for the Chinese listeners might be partly explained by interference of their native tone system. Though they are sensitive to tonal movements, the Chinese tone repertoire established in their perceptual system will constrain the interpretation of the unfamiliar tones.

An issue for further research is the question how important the word tones are for native speakers and for L2 users of Norwegian in speech production and, particularly, speech perception. As Kristoffersen (2000:234) points out, the functional load of the tones is relatively low. In the majority of cases in everyday speech communication, context in a narrow as well as broad sense will supply the listener with sufficient information to make utterances unambiguous. This is also demonstrated by the fact that some dialectal variants of Norwegian lack contrastive word tones. In conclusion, then, we want to maintain that the perception of Norwegian word tones should not receive highest priority in the teaching of Norwegian as a second language. It remains to be investigated to what extent correct production of the tonal melodies might make L2 Norwegian more natural. But then, of course, L2 users would have to master the tones in perception as a prerequisite.

Acknowledgement

The speech material used for the present investigation was developed by Randi A. Nilsen from the Department of Scandinavian Studies and Comparative Literature, Norwegian University of Science and Technology.

References

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-203). Timonium, MD: York Press.

Boersma, P., & Weenink, D. (2007). *Praat: doing phonetics by computer* (Version 4.6.33) [Computer program]. Retrieved 16 October 2007, from http://www.praat.org/.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Timonium, MD: York Press.

Fintoft, K. (1987). Toneme patterns in Norwegian and in Swedish dialects. In R. Channon, & L. Shockey (Eds), *In honor of Ilse Lehiste* (pp. 33-50). The Netherlands: Foris Publications.

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*, 395–421.

Huang, T. (2004). Language-specificity in auditory perception of Chinese tones. Dissertation, Ohio State University.

Kaan, E., Wayland, R., Bao, M., & Barkley, C. M. (2007). Effects of native language and training on lexical tone perception: An event-related potential study. *Brain Research, 1148*, 113-122.

Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, *46*, 295-349.

Kristoffersen, G. (2000). *The phonology of Norwegian*. Oxford, UK: Oxford University Press.

Pruitt, J. S., Jenkins, J.J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *Journal of the Acoustical Society of America*, *119*, 1684-1696.

Rogers, C. L., Dalby, J., & Nishi, K. (2004). Effects of noise and proficiency on intelligibility of Chinese-accented English. *Language and Speech*, *47*, 139-154.

van Dommelen, W. A., & Nilsen, R. A. (2003). Toneme realization in two East Norwegian dialects. Proceedings of *Fonetik 2003, PHONUM, 9* (pp. 21-24). Umeå University, Department of Philosophy and Linguistics.

Van Engen, K., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America*, *121*, 519-526.

Wang, Y., Jongman, A., & Sereno, J. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after training. *Journal of the Acoustical Society of America*, *113*, 1033-1043.

Wang, Y., Jongman, A., & Sereno, J. (2006). Second language acquisition and processing of Mandarin tone. In E. Bates, L. Tan, & O. Tzeng (Eds.), *Handbook of Chinese Psycholinguistics* (pp. 250-257). Cambridge: Cambridge University Press.

Wang, Y., Spence, M., Jongman, A., & Sereno, J. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America, 106*, 3649-3658.

Yudong, C. (2007). From tone to accent: The tonal transfer strategy for Chinese L2 learners. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1645-1648). Saarbrücken, Germany.