Dependence of tone perception on syllable perception

Michael Olsberg [1], *Yi Xu* [1], *Jeremy Green* [2]

[1] Department of Phonetics and Linguistics, University College London, United Kingdom
[2] Orbis Investment Advisory Limited, London, United Kingdom
molsberg@gmail.com, yi.xu@ucl.ac.uk, dplucl@yiremyahu.me.uk

Abstract

We tested the hypothesis that, given the consistency of tone-syllable alignment found in recent research, accuracy of tone perception is dependent on the accuracy of syllable perception. In two experiments, subjects either judged the number of syllables or identified the tones in nonsense sentences that were spectrally intact, low-pass filtered at 300 Hz or converted to sustained schwa carrying the original F_0. It was found that removing spectral information affected not only subjects' ability to judge the number of syllables in a sentence, but also their ability to identify the tones. The results thus confirm the dependence of tone perception on syllable perception.

Index Terms: speech perception, tone perception, syllable perception, tone-syllable alignment

1. Introduction

Although lexical tones in languages like Mandarin are known to be associated with individual syllables, it has never been shown clearly how much tone perception is dependent on syllable perception. In fact, arguments could be easily mounted for the independence of the two in perception. Acoustically, F_0 and spectral property of a sound are largely independent of each other based on the source-filter theory [1]. Articulatorily, F_0 patterns are produced by laryngeal movements, and vocalic and consonantal patterns by supralaryngeal movements, and the two kinds of movements thus should have separate control mechanisms. Phonetically, there have been proposals that tones do not coincide with the syllable, but are carried only by the rhyme or nuclear vowel of the syllable [2, 3]. In phonology, the autosegmental theory asserts that there is full freedom of association of tone with the segmental structure, such that tones are either never lexically associated with a particular syllable or its segmental association can be moved around within a syllable or even across nearby syllables [4].

Recent research, however, has produced accumulating evidence that the production of F_0 and spectral patterns is much more closely related to each other than previously thought. First, in various languages, certain F_0 events, such as peaks and valleys as related to both lexical tone and intonation, are consistently aligned to the edges of the syllable [5-9]. Second, F_0 contours of lexical tones over the interval of an entire syllable in a given tonal context remain largely constant regardless of whether or not the syllable ends with a nasal consonant [10]. Third, such context-dependent contours also remain largely constant regardless of the voicing characteristic of the initial consonant [11]. These findings suggest that there may exist a special mechanism that synchronizes laryngeal and supralaryngeal movements [12].

The observation of the F_0-syllable synchronization raises the possibility that the perception of tones is also dependent on the perception of the syllable. This possibility is especially likely given the manner in which tones are produced as found in [13, 14] and summarized by the Target Approximation (TA) Model [12]. According to the TA model, a tone is produced by articulatorily approaching the underlying pitch target assigned to the syllable. The approximation has to start from an initial laryngeal state, which is actually the final state of the previous syllable, since, due to inertia, the larynx cannot change its state instantaneously [15]. The approximation of the pitch target is continuous throughout the duration of the syllable, although it may asymptote if the syllable is sufficiently long [12]. In other words, despite abrupt shift of underlying pitch targets at each syllable boundary, the surface F_0 is continuous, as can be seen in Figure 1, except when interrupted by voiceless consonants [11].

The continuity of F_0 movements may cause problems for tone perception. To discover an underlying pitch target, the perceptual system needs to know when the target-approaching movement starts and when it ends. To do so, the system needs to segment continuous F_0 patterns into appropriate intervals which should each coincide with a syllable. This could be done by reference to the spectral patterns produced by supralaryngeal articulation. In Figure 1a, for example, the long fall across the second and third syllables should be perceptually divided into movement toward the F tone and that toward the L tone, respectively. Alternatively, tonal segmentation could be done, or at least be assisted by detecting the turning points in the F_0 contours, whenever they are present. In Figure 1, for example, the F_0 contours contain increasing number of turns from top to bottom: 2 in a, 4 in b and 6 in c. If perception is sensitive to the F_0 turning points, tones should be perceived better from tone sequences containing greater number of turns (e.g., c) than those containing fewer number of turns (e.g., a, b).

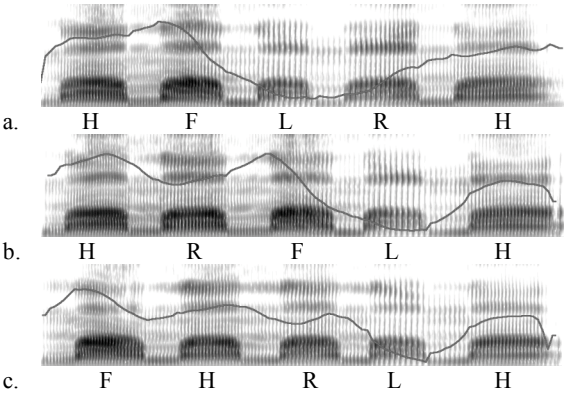

Figure 1: *F_0 contours with varying numbers of turns in the sequence /mamamamama/: 2 turns in a, 4 turns in b, and 6 turns in c. H, R, L and F stand for High, Rising, Low and Falling tones, respectively.*

In the present study, two experiments were carried out to test the hypotheses that a) accurate perception of tone is dependent on accurate perception of the syllable, and b) the perception of tone can be done by detecting the turning points in F_0 contours.

2. Method

2.1. Stimuli

The stimuli consisted of a series of nonsense sentences composed of 3, 4 or 5 /ma/ syllables, as shown in Table 1. The 5-syllable sentences were designed to also have different numbers of turns in their F_0 contours. The sentences with 3 and 5 syllables were recorded by a male native speaker of Mandarin (second author) in the anechoic room of the Department of Phonetics and Linguistics, University College London, using a Bruel & Kjaer sound level meter type 2231 fitted with a 4165 microphone cartridge. Adobe Audition software was used to record the utterances directly onto the hard disk at 44.1 kHz sampling rate with 16 bit quantization. Examples of the recorded sentences are shown in Figure 1.

Table 1. *3-syllalbe and 5-syllable nonsense sentences used in the experiments.*

Tone	Chinese Character	Tone	# of F_0 turns	Chinese Character
FLR	骂马麻	HFLRH	2	妈骂马麻妈
RHF	麻妈骂	LRHFL	2	马麻妈骂马
FRH	骂麻妈	RFRHL	4	麻骂麻妈马
RFL	麻骂马	FRFLH	4	骂麻骂马妈
HRL	妈麻马	FHRLH	6	骂妈麻马妈
LHR	马妈麻	RLHRF	6	麻马妈麻骂

Each sentence had three levels of spectral modifications in order to vary the amount of dynamic spectral information in the signal: no change, low-pass filtering at 300 Hz or replacing the entire spectrum with that of a schwa (hum).

Using Praat, the 4-syllable sentences, which were used only in Experiment 1 for increasing task difficulty, were created by digitally removing either the first or the last syllable of the 5 syllable sentences. This was done by identifying on a spectrogram the beginning of the nasal murmur of the last syllable, or the end of the vocalic section of the first syllable, and digitally removing the material following or preceding it as appropriate.

To eliminate abrupt changes in F_0, due mainly to glottalization, that may be heard as landmarks, the F_0 contour of each sentence was smoothed using Praat's smooth function at a bandwidth of 10 Hz. The smoothed F_0 contours were used to create sustained schwas using Praat's To Sound (hum) function. To reduce inconsistency between the testing conditions, the F_0 contours of the original sentences were also replaced by the smoothed F_0 contours. Finally, the spectrally intact sentences were low-pass filtered at 300 Hz through a Hann pass band to create the filtered sentences.

Due to the unequal loudness of different frequencies [16], a low frequency signal requires a greater intensity than a high frequency signal to be heard as equally loud. Since the peak amplitude of the low-pass filtered sentences were already at the maximum digital value, the amplitude of the unfiltered sentences and the sustained schwas were scaled down by 8 dB, per calculations based on the Equal Loudness Curve [16].

For experiment 1, each sentence was repeated twice, generating 18 sentences x 3 filtering conditions x 2 repetitions = 108 trials. For experiment 2, each sentence was repeated 3 times, generating 12 sentences x 3 filtering conditions x 3 repetitions = 108 trials.

2.2. Subjects

Ten native or near-native speakers of Mandarin, 7 males and 3 females, took part in the two experiments. Their average age was 25.2 (range: 22-31). The average time they lived in Beijing was 17 years (range: 4-26). None of them reported any hearing problems. Nine of them took part in both experiments, while one male subject took part only in experiment 2.

2.3. Experiment 1

2.3.1. Procedures

The subjects carried out the experiment on their own personal computers through a web based interface. Their task was to determine how many syllables they heard, ranging from 1-10 syllables, which was a large enough range to avoid subjects realising that sentences always contained 3, 4 or 5 syllables. Subjects listened to the stimuli through headphones, built-in speakers or loudspeakers. Each trial was heard at least once and they had the option of hearing it two more times.

2.3.2. Analysis and results

The subjects were awarded 1 point for judging the number of syllables correctly and 0 point for an incorrect judgment. Two sets of scores were recorded.

(A) To test the effect of sentence length together with that of level of filtering, a score was awarded to a subject for each length condition (3, 4 and 5 syllables) within each filtering condition (unfiltered, low-pass filtered, hum). Each of the 9 groups contained 6 sentences, each of which was tested twice. Therefore a score of up to 12 was given to a subject for each experimental group. The mean scores are shown in Figure 2.

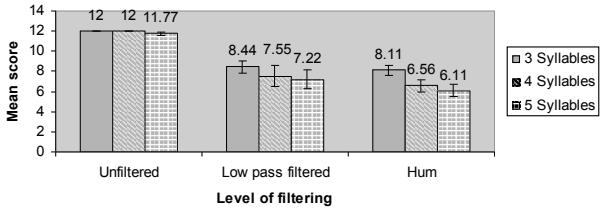

Figure 2: *Mean scores for syllable number judgments, broken down by level of filtering and length of sentence.*

A two-way repeated measures ANOVA showed a significant effect of level of filtering ($F(2,16) = 43$, $p < 0.01$). Bonferroni post-hoc tests showed that while unfiltered sentences scored significantly better than low-pass filtered ones ($p = 0.001$) and hums ($p < 0.001$), there was no significant difference between low-pass filtered sentences and hums. Although in each filtering condition, the longer the sentence, the fewer the number of correct responses, no significant effect of length was found. There was no interaction between sentence length and level of filtering.

(B) To test the effect of number of F_0 turns together with the effect of level of filtering, a score was awarded to a subject for each F_0-turn condition (2, 4 and 6 turns) within each filtering condition (unfiltered, low-pass filtered, hum), for only the 5-syllable sentences. Each condition contained 2 sentences, which were each tested twice. Each subject therefore received a score of up to 4. The mean results are shown in Figure 3.

A two-way repeated measures ANOVA showed a significant effect of filtering ($F(2,16) = 28.4$, $p < 0.001$). Bonferroni

post-hoc tests showed that, while unfiltered sentences scored significantly better than low-pass filtered ones ($p = 0.02$) and hums ($p < 0.01$), there was no significant difference between low-pass filtered sentences and hums. Although the means showed that the more F_0 turns there were in a sentence the higher the score, the effect was not significant. There was also no significant interaction between level of filtering and number of F_0 turns.

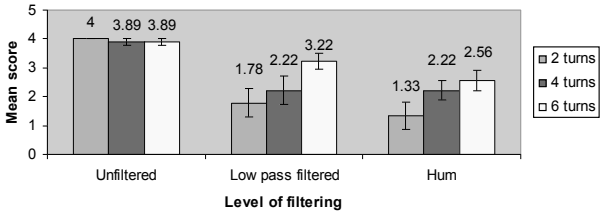

Figure 3: *Mean scores broken down by level of filtering and number of F_0 turns.*

In summary, when higher-frequency dynamic spectral information was removed, either by low-pass filtering, or by full replacement with the static spectrum of a schwa, syllables could not be accurately perceived, despite the presence of F_0 contours.

2.4. Experiment 2

2.4.1. Procedures

Subjects determined the tones of the syllables by selecting from a set of Chinese characters on the screen that differed phonologically only in tone. As in experiment 1, the subjects carried out the experiment using a web based interface using their own personal computers. Each repetition of a sentence was heard at least once and subjects had the option of hearing it two more times. After hearing each sentence, they selected all the syllables they heard. For each syllable heard, they had a choice of four possible characters shown on the screen, each having the H, R, F or L tone. They had the option of entering a choice for up to 10 syllable positions.

2.4.2. Analysis and results

While judging the tones of each sentence, subjects also implicitly judged the number of syllables in the sentence, as they had to decide how many character choices they need to make. When the number of syllables were judged wrongly, it cannot be known which tonal judgment corresponds to which syllable location. Therefore, a scoring system was devised that treated the trials with correct syllable number judgments separately from those with wrong syllable number judgments, as shown in Tables 2 and 3, respectively.

Table 2. *The scoring system if the correct number of syllables were identified.*

	3 syllable sentence	5 syllable sentence
All correct	3 points	5 points
1 incorrect	2 points	4 points
2 incorrect	1 point	3 points
3 incorrect		2 points
4 incorrect		1 point

For example, given the tone sequence HHHHH, an answer of HHHHH would get 5 points, an answer of HHHHL would get 4 points, an answer of HHHLL would get 3 points, and so on.

Table 3. *Scoring system if the incorrect number of syllables was perceived.*

	3 syllable sentence	5 syllable sentence
1 away from total	1 point	2 points
2 away from total	0.5 points	1.5 points
3 away from total		1 point
4 away from total		0.5 points

For example, given a sentence with 5 H tones, an answer of HHHH would get 2 points, as would an answer of HHHHHH. Similarly, an answer of LLLL or LLLLLL, would also get 2 points. This means that subjects would be penalised for judging the wrong number of syllables, but not for misidentifying the tones.

Two sets of data were computed using this scoring system:

A) To test the effect of sentence length together with that of level of filtering, scores were awarded for each length condition within each filtering condition. Each of the 3 groups of the 3-syllable sentences (unfiltered, low-pass filtered, hum) consisted of 6 sentences each of which was tested 3 times. The maximum number of points possible for each group was 54 (3 points for each of 18 sentences). The 5-syllable sentences also had 3 filtering conditions. Each of these 3 groups also consisted of 6 sentences and were tested 3 times. The maximum number of points possible for each group was 90 (5 points for each of the 18 sentences). In order to normalise the results for both length conditions, the score for each group was converted to percentage of the maximum number of points possible. The results are displayed in Figure 4.

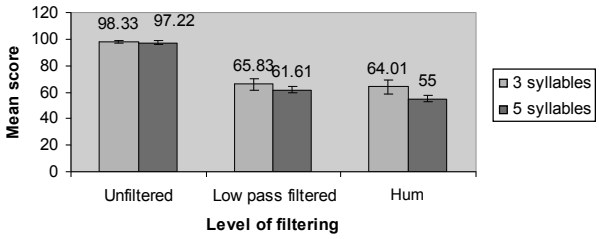

Figure 4: *Mean percentages of correct tone judgment for each length condition within each filtering condition.*

A two-way repeated measures ANOVA showed a significant main effect of level of filtering ($F(2,18) = 123.6$, $p < 0.001$). Bonferroni post-hoc tests showed significantly better performance in the unfiltered condition than in the low-pass filtered ($p < 0.001$) and hum condition ($p < 0.001$). However, there was no significant difference between the low-pass filtered and hum conditions. There was no main effect of length. And there was no interaction between length and level of filtering.

B) To test the effect of number of F_0 turns, the scores were divided according to number of F_0 turns within each filtering condition. Each of the 9 groups of two 5-syllable sentences (3 F_0-turn conditions [2, 4 and 6 turns] and 3 filtering conditions [unfiltered, low-pass filtered and hum]) were tested three times. Thus the maximum score for each group was 30 (5 points for each of the 6 sentences). The mean scores are shown in Figure 5.

A two-way repeated measures ANOVA showed a significant main effect of level of filtering ($F(2,18) = 123.6$, $p < 0.001$). Bonferroni post-hoc tests showed significantly better performance in the unfiltered condition than in the low-

pass filtered ($p < 0.001$) and hum ($p < 0.001$) conditions. The difference between the low-pass filtered and hum conditions was not significant. There was no main effect of number of F_0 turns. And there was no interaction between number of F_0 turns and level of filtering.

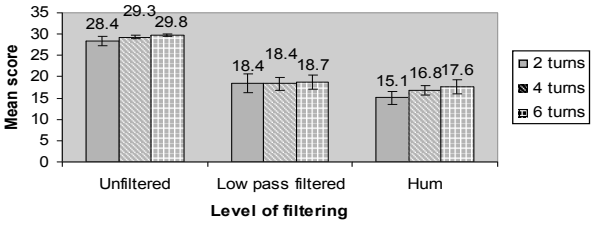

Figure 5: *Mean score for each number of F_0 turns within each level of filtering.*

In summary, when higher-frequency dynamic spectral information was removed, either by low-pass filtering, or by full replacement with the static spectrum of a schwa, tones could not be accurately perceived from the F_0 contours alone.

3. Discussion and Conclusion

Experiment 1 shows that native speakers of Mandarin require dynamic higher-frequency spectral information in order to accurately perceive syllables in an utterance. Syllables that had been low-pass filtered at 300 Hz as well as those whose entire spectra have been replaced with that of a schwa were much more poorly perceived than the spectrally intact syllables. The perceptual rates did not vary with either sentence length or number of F_0 turns. In the materials used in the present study, which were nonsense sentences consisting of the syllable [ma], the only salient spectral information that remained after low-pass filtering at 300 Hz was the nasal murmur, as can be inferred from Figure 1. The movement of none of the formants could be effectively transmitted, given that the first formant of the [a] was well above 300 Hz. This suggests that in order for syllables to be accurately perceived, listeners need to hear the formant movements. It would be a question for future research how much more formant information is needed to accurately perceive the syllables.

Experiment 2 shows that when accurate syllable information is not available, as demonstrated by Experiment 1, tones cannot be accurately perceived either. It also shows that turning points in F_0 contours alone cannot provide accurate information about the identity of the tones when dynamic spectral information is not present. This is despite a slight, but nonsignificant, trend in the direction that more F_0 turns leads to more accurate judgments in both experiments.

The results of the two experiments therefore support the hypothesis that tone perception in Mandarin is inextricably tied to that of the syllable. This finding compliments the previous finding that the production of tones is inextricably tied to that of the syllable [10, 12, 13, 14], and that F_0 turning points are byproducts of realizing tones through articulatory target approximation rather than being the direct correlates of tones [17].

Given that consistent F_0-syllable alignment similar to that of Mandarin has also been found in a number of non-tone languages, including Greek [5], English [6], Dutch [7], German [8], and Italian [9], it is possible that syllable-synchronized target approximation is also the basic mechanism of pitch production in those and probably other non-tone languages. If so, the perception of intonational pitch events in those languages would also be dependent on the perception of the syllable. This possibility could be tested in future research using similar methods as used in the present study.

Finally, the present finding may also benefit automatic speech recognition, as it demonstrates the critical importance of exact F_0-syllable alignment for the recognition of tones.

4. Acknowledgements

This work is supported in part by NIH grant 1R01DC006243 to the second author.

5. References

[1] Fant, G., Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

[2] Howie, J. M., "On the domain of tone in Mandarin", Phonetica, Vol. 30, 129-48, 1974.

[3] Lin, M., "A perceptual study on the domain of tones in Standard Chinese", Chinese Journal of Acoustics, Vol. 14, 350-57, 1995.

[4] Goldsmith, J.A., Autosegmental and Metrical Phonology, Blackwell Publishers, Oxford, 1990.

[5] Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. "What is a starred tone? Evidence from Greek." In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, edited by Michael B. Broe and Janet B. Pierrehumbert, 119-31. Cambridge: Cambridge University Press, 2000.

[6] Ladd, D.R., D. Faulkner, H. Faulkner, and A. Schepman, "Constant "segmental anchoring" of F0 movements under changes in speech rate", J. Acoust. Soc. Amer., Vol. 106, 1543-54, 1999.

[7] Ladd, D. R., I. Mennen, and A. Schepman, "Phonological conditioning of peak alignment in rising pitch accents in Dutch", J. Acoust. Soc. Amer., Vol. 107, 2685-96, 2000.

[8] Atterer, M., and D. R. Ladd, "On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German", J. Phon., Vol. 32, 177-97, 2004.

[9] D'Imperio, M., R. Espesser, H. Lœvenbruck, C. Menezes, N. Nguyen, and P. Welby. "Are tones aligned with articulatory events? Evidence from Italian and French." In *Papers in Laboratory Phonology IX: Change in Phonology*, edited by J. Cole and J. Hualde. The Hague: Mouton de Gruyter, in press.

[10] Xu, Y., "Consistency of tone-syllable alignment across different syllable structures and speaking rates", Phonetica, Vol. 55, 179-203, 1998.

[11] Xu, C. X., and Y. Xu, "Effects of consonant aspiration on Mandarin tones", Journal of the International Phonetic Association, Vol. 33, 165-81, 2003.

[12] Xu, Y., and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese", Speech Communication, Vol. 33, 319-37, 2001.

[13] Xu, Y., "Contextual tonal variations in Mandarin", J. Phon., Vol. 25, 61-83, 1997.

[14] Xu, Y., "Effects of tone and focus on the formation and alignment of f_0 contours", J. Phon., Vol. 27, 55-105, 1999.

[15] Xu, Y., and X. Sun, "Maximum speed of pitch change and how it may relate to speech", J. Acoust. Soc. Amer., Vol. 111, 1399-413, 2002.

[16] Fletcher, Harvey, and W. A. Munson, "Loudness, Its Definition, Measurement and Calculation", J. Acoust. Soc. Amer., Vol. 5, 82-108, 1933.

[17] Xu, Y., "Speech melody as articulatorily implemented communicative functions", Speech Communication, Vol. 46, 220-51, 2005.