Revisiting pitch slope and height effects on perceived duration

Carlos Gussenhoven[1], *Wencui Zhou*[2]
[1]Radboud University Nijmegen, Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
c.gussenhoven@let.ru.nl, wencui.zhou@mpi.nl

Abstract

The shape of pitch contours has been shown to have an effect on the perceived duration of vowels. For instance, vowels with high level pitch and vowels with falling contours sound longer than vowels with low level pitch. Depending on whether the comparison is between level pitches or between level and dynamic contours, these findings have been interpreted in two ways. For inter-level comparisons, where the duration results are the reverse of production results, a hypercorrection strategy in production has been proposed [1]. By contrast, for comparisons between level pitches and dynamic contours, the longer production data for dynamic contours have been held responsible. We report an experiment with Dutch and Chinese listeners which aimed to show that production data and perception data are each other's opposites for high, low, falling and rising contours. We explain the results, which are consistent with earlier findings, in terms of the compensatory listening strategy of [2], arguing that the perception effects are due to a perceptual compensation of articulatory strategies and constraints, rather than that differences in production compensate for psycho-acoustic perception effects.

Index Terms: perception-articulation interaction, pitch contour, perceived duration, hypercorrection strategy, compensation strategy

1. Introduction

The shape of the fundamental frequency (f_0) contour affects the perception of the duration of the vowel on which it is located [1], [3], [4]. Broadly, two findings have been reported. One is that vowels with complex dynamic pitch contours, i.e., rising-falling and a falling-rising contours, are perceived as longer than level pitch [3], [4], and another that the perceived duration of vowels correlates with the height of level pitch, as measured over three heights [1]. Comparisons between the effects of non-complex dynamic contours, i.e. a fall or a rise, and those of level pitch were ambiguous in the investigation in [1], since vowels with a rise and fall were perceived as longer than those with low level pitch, but only vowels with a fall were heard as longer than mid level pitch. Vowels with a rise were perceived as shorter than mid-pitch vowels, while no significant effect was found for high level pitch in comparisons with rises and falls.

When these findings are related to production data, they appear to fall into two groups. Comparisons between the perceived duration scores for complex dynamic contours and level contours would appear to replicate production differences, since vowels with fall-rises and rise-falls take longer to produce than vowels with level tones. By contrast, comparisons among the perceived durations of vowels with level pitch result in a reversal of the differences between acoustic durations. Comparisons of the direction of the differences between perceived and acoustic durations in the case of simple dynamic contours vs. level pitch produced ambiguous results. A provisional explanation for the reversal of the relation between acoustic and perceived durations was presented in [1]. This account relies on a compensatory articulation policy intended to undo the perception effects, characterized as a hypercorrection effect. In this scenario, the perception effects themselves will most probably have a non-linguistic, psycho-acoustic source, an assumption which is supported by the results of [5] that (non-vowellike) high level tones sound longer than low level tones, all else being equal.

The purpose of the experiment reported here was twofold. First, we intended to present data that would allow us to argue that the explanation for the reversal of the effects of pitch slope and height has a basis in a general compensatory mechanism that operates in the perception of linguistic stimuli. That is, instead of assuming that the compensatory behaviour occurs during the production of vowels with various pitch contours, the assumption here is that the compensation occurs during perception [2]. Generally, listeners may fail to include side effects of the articulation of a primary phonetic parameter on some secondary phonetic parameter in the perceived value of the secondary parameter (cf. [9]). For instance, since physiological conditions favour higher f_0 during the production of high vowels like [i] than during the production of lower vowels like [a], a side effect of tongue height on f_0, listeners will deduct some portion from the acoustic f_0 before assigning it a value for perceived f_0 (pitch), causing high vowels to have lower pitch than low vowels when their f_0 specifications are identical [6], [7]. Similarly, since speech at the beginning of a breath group tends to have higher f_0 than speech later in the breath group, listeners will deduct some of the f_0 of earlier f_0 peaks as compared with later f_0 peaks, causing earlier peaks to have lower pitch [8]. A third effect of this kind was reported by [2], where perceived duration was shown to depend on vowel height. For anatomical reasons, high vowels take less time to produce than low vowels. If listeners apply the policy of deducting some of the acoustic duration of lower vowels as compared to higher vowels, the prediction is that higher vowels sound longer than lower vowels, an effect that was found for Dutch listeners.

The second aim was to extend the class of contours for which the perceived duration mirrors the durations of the articulations to simple dynamic contours, falls and rises. This extension does not in fact go against the results of [1], which are neutral with respect to whether perceived durations follow production data or are the reverse of production data. There are two implications of our position. The first is that differences in the perceived duration of vowels with different pitch patterns do not necessarily apply to non-linguistic stimuli. Second, durational differences in articulation must be independent, and thus require an explanation that relies on constraints or strategies inherent in the production of f_0 patterns. The first implication is left for future research; the second will be taken up in the conclusions.

2. Method

2.1. Stimuli

In order to maximize linguistic realism, we decided not to use isolated vowels as the source utterances of our stimuli, but disyllabic structures in which the target vowel is embedded in the second syllable. Because we wanted to emphasize the language independent nature of the effects, we ran the perception experiment with participants who had either Standard Dutch or Standard Chinese as their first language and obtained the source utterances for the stimuli from a native speaker of Russian, so as to avoid a participant language bias.

We recorded the disyllables [mek^{h}a], [meka], [mek^{h}i], and [meki] as produced by a phonetically trained female Russian speaker. After splicing off the parts of the speech signals starting at the first vocal pulse of the last vowel, we selected one of the spliced off parts that represented [a] and one that represented [i], such that out of the four possible pairings, their durations and intensities were most alike ([a] 199 ms, [i] 197 ms, [a] 57.67 dB, [i] 64.38 dB). These sound files were subjected to duration and pitch manipulations, using Praat [10]. First, a 4-step duration continuum was made with durations of 230, 260, 290 and 320 ms. These steps are considerably smaller than those used by other researchers, but these values are approximately located within a window found in natural languages, following our objective to have realistic linguistic stimuli. Each step for both vowels was provided with six f_0 patterns, high level (HH), low level (LL), high rising (LH), high falling (HL), high curving (LHL), and low curving (HLH). Table 1 provides the f_0 specifications of these pitch contours, chosen such that they broadly matched the voice of our female speaker. The 24 versions of [a] were spliced onto the remainders of the two speech files that had [a] as the final vowel, while the equivalent versions of [i] were spliced onto the remainders of the two speech files that had [i] as the final vowel. We normalized the durations of the post-burst sections of the speech files, the VOTs, by taking the average of the original values (37 ms in the case of [meka] and [meki] and 87 ms in the case [mek^{h}a] and [mek^{h}i]), but left the preceding portions of the speech files intact, with total durations of 458 ms for [mek(^{h}a)], 439 ms for [mek(^{h}i)], 437 ms for [mek(a)] and 445 ms for [mek(i)].

Table 1. *Pitch values of the six contours used in the present study.*

Pitch contour	Pitch beginning (Hz)	Pitch middle (Hz)	Pitch ending (Hz)
HH	250	250	250
LL	140	140	140
LH	140	207.5	275
HL	275	207.5	140
LHL	140	275	140
HLH	275	140	275

2.2. Participants

Twenty native Mandarin speakers (average age = 24) and twenty native Dutch speakers (average age = 23) participated in the experiment. Both groups of participants were undergraduate and graduate students recruited from Radboud University Nijmegen in the Netherlands. None of the Dutch participants had any prior experience with any tone language. All participants reported to have normal hearing.

2.3. Procedures

Stimuli were presented twice, using an experiment file running in Praat [10]. The 192 trials were randomized per participant and presented in an AX task in which "A" was used as an anchor stimulus and "X" as the target stimulus. Participants were told they were going to listen to pairs of pronunciations of a word from a foreign language and were asked to judge the duration of the last vowel of the second pronunciation. They were explicitly instructed to make their judgements in comparison with the equivalent vowel in the first pronunciation, and the register the degree to which they judged the second vowel as either shorter or longer than the first on a 7-point scale, with 1 being the shortest and 7 the longest. The anchor stimulus represented an average stimulus, with a second vowel having level pitch of 195 Hz and a duration of 275 ms. They were allowed to listen to the stimuli as often as they needed before making their judgment.

Before the experiment started, participants were given three practice trials. In one trial, the target vowel had an average duration (275 ms), as in the anchor stimulus. In the other two trials, the target vowel had either the shortest duration or the longest duration. After these practice trials, participants completed the 192 test trials without pause. The experiment was conducted in a soundproof booth at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. Participants were seated in front of an 15.5-inch HP laptop, wearing high quality headphones (Sennheiser HD 201). They were allowed to adjust the volume to their preference. The total duration of the experiment was 25 minutes. They were paid a small fee.

3. Results

Perception scores were analyzed with Repeated Measures Analyses of Variance. In the analysis, LANGUAGE GROUP was included as a between-subjects factor, and CONSONANT ([k^{h}], [k]), VOWEL ([a], [i]), DURATION STEP (230, 260, 290 and 320 ms) and CONTOUR (HH, LL, LH, HL, LHL, HLH) as within-subjects variables. Mauchly's Test for Shpericity was significant for both DURATION STEP and CONTOUR, and we report Greenhouse-Geisser corrected significance levels only. The analysis showed significant main effects of CONSONANT, VOWEL, DURATION STEP and CONTOUR (Table 2), and significant two-way interactions CONSONANT × LANGUAGE GROUP, VOWEL × LANGUAGE GROUP, VOWEL × DURATION STEP, CONTOUR × DURATION STEP, and VOWEL × CONTOUR (Table 3). There was no main effect of LANGUAGE GROUP. There was a three-way interaction between LANGUAGE GROUP, VOWEL and DURATION [$F(1, 18) = 46.51$, $p < .001$].

Table 2. *Main effects of consonant, vowel, duration step and contour.*

Variables	Results
Consonant	$F(1, 38) = 183.17$, $p = .001$
Vowel	$F(1, 38) = 46.51$, $p < .001$
Duration step	$F(1.29, 48.96) = 290.51$, $p < .001$
Contour	$F(3.30, 125.55) = 11.88$, $p < .001$

Table 3. *Two-way interactions*

Interactions	Results
Consonant*Language	$F(1, 38) = 14.14$, $p < .01$
Vowel*Language	$F(1, 38) = 4.36$, $p < .05$
Vowel*DurationStep	$F(2.56, 97.84) = 10.61$, $p < .001$
Contour*DurationStep	$F(9.30, 353.52) = 4.31$, $p < .001$
Vowel*Contour	$F(4.59, 174.30) = 6.46$, $p < .001$

The effect of CONSONANT was due to the longer perceived duration of the vowel after [k^h] than after [k]. The interaction with LANGUAGE is due to the stronger effect in the Chinese group than in the Dutch group. This is shown in Fig. 1. The effect of VOWEL was likewise due to a longer perceived duration of [a] than of [i], an effect that was stronger in the Chinese group than in the Dutch group, which caused the interaction with LANGUAGE. However, while the Chinese group differentiates the two vowels at all four durations, the Dutch group only does so at the longer duration steps, as shown in Fig. 2. This is the cause of both the two-way VOWEL × DURATION STEP and the three-way VOWEL × DURATION STEP × LANGUAGE interactions. This result is shown in Figure 2.

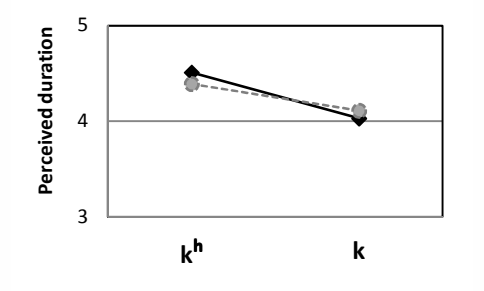

Figure 1: *Interaction between CONSONANT and LANGUAGE (___= Chinese, _ _ =Dutch).*

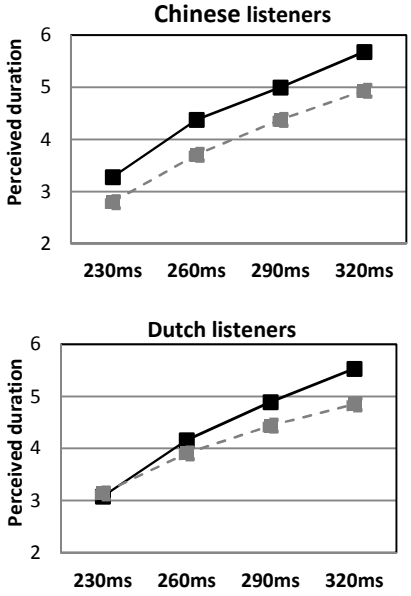

Figure 2: *Interaction between VOWEL and DURATION STEP (___= [a], _ _ =[i]), for Chinese and Dutch listeners separately.*

The effect of DURATION STEP is as expected. Each 30 ms increase in acoustic duration of the target vowel corresponded to a significant increase in perceived duration, according to post-hoc pairwise comparisons ($p<0.001$ for all three steps, Bonferroni corrected). The main effect of CONTOUR is due to a shorter perceived duration for the LL-contour as compared with the other five contours (HH, LH, HL, LHL and HLH). The interaction between CONTOUR and DURATION STEP is due an increasing effect size of the shorter perceived duration of LL at the higher duration steps. This is shown in Fig. 3. To some extent the greater differential relief among the contours is also seen in the longer perceived duration of HL in the highest duration step.

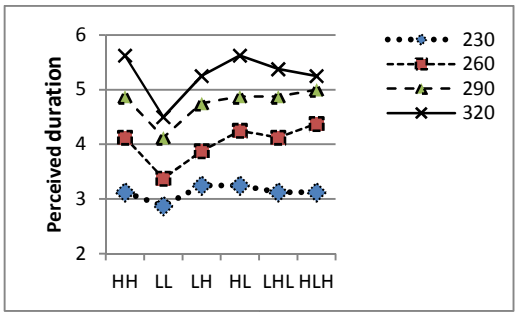

Figure 3: *Interaction between CONTOUR and DURATION step.*

Finally, the VOWEL × CONTOUR interaction is due to the greater differentiation of the levels of CONTOUR for the stimuli with [a] than for those with [i]. Figure 4 presents this finding.

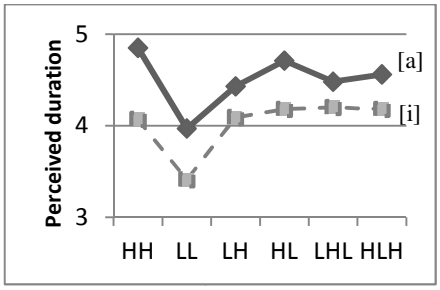

Figure 4: *Interaction between VOWEL and CONTOUR (___= [a], _ _ =[i]).*

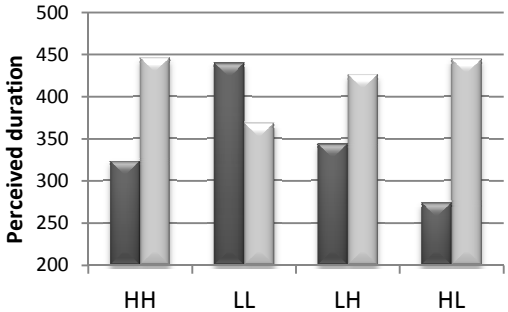

Figure 5. *Comparison of production data from [11] (black) with perception data obtained in the experiment reported here (grey).*

In addition to the Analysis of Variance we addressed the issue of the relation between perceived and produced durations more directly by calculating the Pearson correlation coefficient between our data and production data presented by [11] for the four lexical tones of Standard Chinese. These are averaged values over six pronunciations of the syllable [ba], figuring in four monosyllabic words with each of the four lexical tones, taken from their Table 1. This gave an $r = -0.95$, meaning that 90% of the variance in perceived duration is explained by independently obtained production data. This negative correlation is shown in Fig. 5.

4. Discussion

In addition to the findings directly related to the research question, we have established that listeners hear the final vowel in a disyllabic structure as longer if the onset consonant has a 50 ms longer VOT (Fig. 1), an effect that was slightly stronger for the Chinese than for the Dutch participants. Since the vowel duration in the stimuli was defined as the voiced portion of the vowel and thus excluded the VOT, this suggests that listeners include the VOT as part of the duration of the vowel. The effect sizes of the VOT difference and the duration steps are in fact closely commensurate. Extrapolating from the scores for the duration steps, a difference of 50 ms should give a difference in perceived duration of 0.403 scale points; the difference for the consonants is 0.380 scale points, averaged over the two listener groups. The explanation of the larger effect in the Chinese group must be the phonological status of late VOT in Standard Chinese, where aspirated voiceless plosives contrast with unaspirated ones; in Dutch, prevoiced plosives contrast with voiceless unaspirated ones.

The unpredicted effect of the nature of the vowel is harder to account for, as is the fact that the effect is stronger in the Chinese group. It should be borne in mind that the purpose of the vowel variable was to increase the variation in the stimuli and no effort was taken to normalize intensity envelopes or intensity levels. What our results do suggest, however, is that [a] may provide a better acoustic background for the compensation effects we have found than [i] (Fig. 2 [Dutch] and Fig. 4), something that was also suggested by the fact that researchers have generally chosen this vowel for their stimuli.

Against the background of our research question, the more interesting finding is the high negative correlation between acoustic durations obtained from a production study [11] and the perceived durations for corresponding f_0 contours presented in our stimuli. The perceptual compensation theory makes precisely this prediction, while predictions about differences between perceived durations of one f_0 contour and the next are derivative, and should be evaluated against the size of the difference in production. The results from the Analysis of Variance in fact perfectly replicate equivalent results in [1] and neither investigation thus produced all the results that could have been predicted by our theory. Specifically, while falls have generally been found to be shorter than rises (for references, see [1]), the perceived duration of simple falls is not significantly longer than that of simple rises. Here, the effect size is probably too small for that result to arise in these small-scale perception experiments.

As said at the end of section 1, our perceptual compensation theory presupposes independent explanations for durational differences in the production of f_0 contours. The shorter duration of falls than rises has been widely discussed and investigated since [12]. No explanation has been proposed for the long duration of low tones than that of high tones. We believe that the longer duration of low tones serves as an enhancement, along with flanking f_0 movements towards and away from the low pitch target. Mandarin Tone 3 is well known for being a low 'dipping' contour, with a preceding fall and - in phrase-final position - with a final upward slope. English L*-tones were described as 'dips' by [13], while the L* of Stockholm Swedish Accent 1 shows variable presence of a leading H-tone [14]. The extra duration, while partly a consequence of the dipping movement, may independently contribute to the enhancement. High tones by contrast will often just need to reach high f_0 so as to be parsed as H-tones.

A final comment is made on the results for the complex dynamic f_0 contours, the rising-falling LHL and the falling-rising HLH. Perceptual compensation predicts that these contours should be perceived as shorter than the simple dynamic contours, since their greater complexity predicts longer duration. The fact that they are not need not mean that they are not subject to any compensation strategy, since it could be the case that the perpetual compensation has its limits. However, we can at least conclude is that the equal perceived duration for simple and complex dynamic f_0 contours does not confirm any theory predicting that perceived durations are biased in the direction of acoustic durations in production.

5. Conclusion

Our investigation and the interpretations of the results allow for two conclusions.

First, listeners compensate for side effects of articulations. The perceived value of the phonetic parameter affected by an articulatory side effect may have undone the side effect, with the result that perceived values are the reverse of acoustic values in articulation. This perceptual compensation theory explains why the pitch of low vowels is higher than that of high vowels, where the side effect is the intrinsic f_0; why high vowels sound longer than mid vowels, where the side effect is that of the tongue-palate distance on duration; why early f_0 peaks have lower pitch than later f_0 peaks, where the side effect is declination; and, as argued in this contribution, it explains why low tones are perceived as shorter than high tones, while predicting that rises are heard as shorter than falls. Here, the side effects are due to a greater difficulty of producing rising f_0 as compared to falling f_0 and a strategy of enhancing low tones.

Second, listeners judge the duration of a post-stop vowel as beginning at the release of the closure, and it thus includes the (positive) VOT. This may explain the interactions between adjustments in the duration of the voiced portion and the VOT reported for English by [15]. If durational adjustments target zones in the interval defined by the VOT and the voiced portion of the vowel, the proportions by which the VOT and the voiced portion are affected will vary with the location of the zone (e.g. initial half or final half).

6. Acknowledgements

We thank Joop Kerkhoff for his help with the stimulus preparation and the experiment set-up, Mirjam Broersma for statistical support and Tomas Riad for the quick way in which he provided some useful information.

7. References

[1] Yu, A. C. L., "Tonal effects on perceived vowel duration", in C. Fougeron, B., Kuehnert, M., D'Imperio, M. and N. Vallée [Eds], Laboratory Phonology 10, 151-168, Mouton de Gruyter, 2010.

[2] Gussenhoven, C., "A vowel height split explained: Compensatory listening and speaker control", in J. Cole and J.I. Hualde [Eds], Laboratory Phonology 9, 145-172, Mouton de Gruyter, 2007.

[3] Lehiste, I., "Influence of fundamental frequency pattern on the perception of duration", Journal of Phonetics 4, 113–117, 1976.

[4] Pisoni, D.B., "Fundamental frequency and perceived vowel duration", Journal of the Acoustical Society of America 59, S39.

[5] Brigner, W., "Perceived duration as a function of pitch", Perceptual and Motor Skills 67, 301-302.

[6] Hombert, J.-M., "Consonant types, vowel quality, and tone", in V. Fromkin [Ed], Tone: A Linguistic Survey, 77-112, Academic Press, 1978.

[7] Silverman, K., The Structure and Processing of Fundamental Frequency Contours, PhD Dissertation, University of Cambridge, 1987.

[8] Pierrehumbert, J.B., "The Perception of fundamental frequency declination", Journal of the Acoustical Society of America 66, 363-369.

[9] Stevens, K.N. and Keyser, S.J., "Primary features and their enhancement in consonants", Language 65, 81-106, 2009.

[10] Boersma, P. and Weenink, D., Praat: Doing Phonetics by Computer. www.org.nl, 1999-2012.

[11] Whalen, D. and Xu, Y. "Information for Mandarin tones in the amplitude contour and in brief segments", Phonetica 49, 25-47, 1992.

[12] Ohala, J.J. and Ewan, W.G., "Speed of pitch change", Journal of the Acoustical Society of America 53, 345-345.

[13] Leben, W.R., "The tones of English intonation", Linguistic Analysis 2, 67–107, 1976.

[14] Engstrand, O., "Phonetic interpretation of the word accent contrast in Swedish: Evidence in spontaneous speech", Phonetica 54, 61-75, 1997.

[15] Port, R.F. and Rotunno, M., "Relation between voice-onset time and vowel duration", Journal of the Acoustical Society of America 66, 654-662, 1979.