PITCH, PERCEIVED DURATION AND AUDITORY BIASES: COMPARISON AMONG LANGUAGES

Juraj Šimko[1], Daniel Aalto[2], Pärtel Lippus[3], Marcin Włodarczak[4], Martti Vainio[1]

[1]University of Helsinki, Finland, [2]University of Alberta, Canada, [3]University of Tartu, Estonia,

[4]Stockholm University, Sweden

juraj.simko@helsinki.fi, aalto@ualberta.ca, partel.lippus@ut.ee, wlodarczak@ling.su.se, martti.vainio@helsinki.fi

ABSTRACT

In addition to fundamental frequency height, its movement is also generally assumed to lengthen the perceived duration of syllable-like sounds. The lengthening effect has been observed for some languages (US English, French, Swiss German, Japanese) but reported to be absent for another (Thai, Latin American Spanish, German). In this work, native speakers of Estonian, Finnish, Mandarin and Swedish performed a two-alternative forced choice duration discrimination experiment with pairs of complex tones varying in several acoustic dimensions. According to a logistic regression analysis, the duration judgements are affected by intensity, f_0 level, and f_0 movement for all languages, but the strength of these influences varies across languages and a pattern revealed by the relative strengths correlates with phonological properties of the languages. The findings are discussed in the light of current hypotheses of the origin of pitch modulation of perceived duration.

Keywords: perceived duration, fundamental frequency, psychoacoustics, hyper-correction, auditory phonetics

1. INTRODUCTION

It has been a long established observation in phonetic and psychoacoustic literature that, when comparing duration of two sounds, listeners tend to judge the sound with a higher f_0 or dynamic pattern as longer than a sound with a lower or static f_0 contour of the same physical duration.

While f_0-modulation of perceived duration for static sounds – higher sounds judged as longer compared to lower ones – has been robustly supported for listeners from various language backgrounds the relation between f_0 dynamicity and perceived duration seems to be more complex [2, 13, 21, 1, 26, 9, 28, 4, 7]. Earlier investigations involving US English found that, in general, tones with non-static pitch patterns were judged longer than the those with level f_0 [13, 20, 27, 28]. These results also suggested a degree of correlation between perceived duration and the extent and direction of f_0 movement: a greater movement results in increased perceived duration with rises judged as longer than falls [21].

Subsequent studies revealed substantial language dependency on listeners' behavior. Lehnert-LeHouillier [14] compared the influence of the lengthening effect on speakers of four languages without (Latin American Spanish) and with (German, Thai and Japanese) phonological vowel length contrast. Moreover, unlike in German and tonal Thai, in Japanese the contrast is co-signaled by greater pitch movement (fall) during the phonologically long vowels [10, 23]. The results showed that only Japanese participants judged stimuli with falling pitch as longer compared to level f_0 contour.

On the other hand, a recent study by Cumming [4] involving speakers of Swiss German, Swiss French and French found a robust lengthening effect of pitch movement on perceived duration of both speech and non-speech stimuli. The analysis has not revealed any difference between the languages.

Finally, Gussenhoven [7] compared the duration-modulation effect of f_0 level and tone shape on Dutch and Mandarin participants, and found significant quantitative differences based on language background. Interestingly, the order in which different tonal shapes influenced durational judgements negatively correlated with the known durational patterns in production: the sounds that are produced shorter in tonal languages [6, 11] (HH, LH) were judged as *longer* than their phonetically longer counterparts (LL, HL) (see also [28]).

These results bring forth a question of the origin and universality of the pitch-modulation of duration. The proposed explanations usually involve the hyper-correction hypothesis [19]. Its "auditory-based" version states that pitch-modulation is grounded in the properties of human auditory apparatus and leads to compensations in production

patterns driven by economy requirements [28, 15]. A "production-based" counterpart argues for a predominance of articulatory influences: some sounds, e.g. HH, are produced shorted due to physiological constraints; listeners subsequently generalize this phenomenon and overestimate the duration of such sounds in their perceptual *judgements* [7]. The assumption of auditory grounding of modulation effects is in fact dispensable for the latter account.

We aim to contribute to this discussion by investigating pitch-modulation of duration for four languages: Estonian, Finnish, Mandarin and Swedish. The language selection is based on the way in which pitch and duration interact in their phonological systems and phonetic realization. Mandarin is a tone language with no quantity contrast; its speakers use primarily f_0 to mark the contrast [18]. On the other hand, Estonian and Finnish are quantity languages that use pitch movement alongside duration to co-signal quantity contrast [16, 24, 17, 22, 25, 8]. In contrast, Swedish has a complementary quantity system in stressed syllables which is manifested primarily by duration and, to a lesser degree, by vowel quality; pitch is not included in standard description of the Swedish quantity system.

Our approach differs from the studies discussed above in two important aspects. First, rather than looking merely for presence or absence of the phenomena under investigation, we evaluate the *strength* of influence of continuously varied f_0 level, its dynamic range as well as actual duration and intensity of sound on duration judgment. The individual effect strengths are then compared between groups based on language background.

As we want to isolate the effects of physical properties of the sound on duration modulation, a great attention was paid to stimulus design. We use non-speech stimuli approximating syllables in duration, pitch level and slope. However, the spectral properties of the stimuli differ from speech considerably: in order to mitigate the physical correlation between frequency and energy, the stimuli are band-passed filtered using a narrow band around 3 kHz.

2. METHODS

A two alternative forced choice duration discrimination task with 400 pairs of stimuli was presented to native speakers of Estonian (N=18), Finnish (N=15), Mandarin (N=15) and Swedish (N=6). The participants had no music background (at most two years of weekly musical activity) and no hearing problems. They heard a pair of sounds through level calibrated headphones and chose which of the stimuli was longer using designated keys on a keyboard. The participants were told to concentrate on the duration and neglect any other variation in the stimuli.

Duration of stimuli was drawn from truncated normal distribution with a mean of 300 ms, a standard deviation of 75 ms, and inclusion interval 150–450 ms. The f_0 level was randomly chosen with a mean of 150 Hz and a standard deviation of 4 semitones. Moreover, an f_0 rise/fall was superimposed with random interval of a mean of 0 and a standard deviation of 4 semitones over stimulus duration. The level of stimuli was drawn from distribution with the mean 66 dB (SPL), standard deviation 1 dB and inclusion threshold of 2 dB were included.

The onsets of the sounds had a random difference of mean 800 ms and standard deviation 10 ms with at most 20 ms deviation from the mean. The sound signals were constructed based on simple sawtooth waves, gamma filtered with center frequency 3141.59 Hz [3]. Before adjusting for the desired level, the intensity of the signals over the first 100 ms was made equal. Finally, the stimuli were masked by white (broadband) noise with 10 dB SNR with respect to the standard sound.

A mixed effects logistic regression model was fitted to the data. The dependent variable was the binary response, the fixed factors were duration difference between the first and second stimulus, level difference, f_0 level difference , difference in f_0 movement (Δf_0, to investigate the effect of pitch movement direction) and difference in absolute value of f_0 movement ($|\Delta f_0|$, to investigate the effect of pitch movement extent). In addition, the interaction terms between all these acoustic factors and language were included as fixed effects. Since individual variation in duration discrimination is large, the (simple) fixed factors were also included as random slopes for subjects which were treated as random effects.

3. RESULTS

Table 1: *Summary of the logistic regression.*

	EST	FIN	SWE	MAN
interc.	0.20	0.18	0.25	0.56^{***}
dur. dif.	20.7^{***}	25.2^{***}	19.0^{***}	16.4^{***}
f_0 dif.	0.17^{***}	0.12^{***}	0.11^{***}	0.07^{***}
Δf_0 dif.	0.05^{***}	0.04^{***}	0.03^{*}	0.03^{***}
$\lvert\Delta f_0\rvert$ dif.	-0.01	0.05^{**}	0.02	0.06^{***}
level dif.	0.15^{***}	0.09^{**}	0.09	0.07^{*}

Tab. 1 summarizes the fit expressed by the model. It shows the regression coefficients for each independent variable per language, as well as statistical

significance of the coefficient being different from zero. The full mixed logistic regression model reduces the deviance by 32.3% compared to a null model. Of this reduction, 85.3 % is accounted for by the duration difference alone while the other factors, level (0.6 %), f_0 level (12.4 %), Δf_0 (1.1 %), $|\Delta f_0|$ (0.6 %), account for only small amount.

The relative impact of the acoustical dimensions on the duration judgments can be seen by comparing the effects against the duration difference term: e.g, for Estonians, the first sound is judged in average 9.4 ms longer, a 1 dB increase in level corresponds to 7.4 ms increase of judged duration, a 1 semitone f_0 level increase lengthens it by 8.4 ms, and a 1 semitone larger f_0 rise (or smaller f_0 fall) by 2.2 ms.

The positive intercepts for all language-based groups show a tendency to judge the first stimulus as longer, everything else being equal. This trend, typical for this type of forced-choice duration comparison experiments [13, 21], is significant only for the MAN group. The coefficient for MAN is significantly different from those for EST and FIN groups ($p < 0.05$); other differences are not significant.

Statistically significant coefficients for duration difference mean that all four groups of subjects responded to the task appropriately, on average judging the stimulus with greater duration as longer. The coefficient sizes for both EST and FIN groups were significantly greater than for MAN ($p < 0.01$), i.e., everything else being equal, the Finns and Estonians made more precise durational judgements than Mandarin speakers.

The judgement of all speaker groups was significantly influenced by pitch level of the stimulus (f_0 difference coefficient). As the coefficient sizes indicate, the influence was greatest for Estonians, followed by Finnish, Swedish and Mandarin speakers. The following differences in effect size are statistically significant: EST>MAN ($p < 0.001$), EST>FIN and FIN>MAN ($p < 0.01$), EST>SWE ($p < 0.05$).

The Δf_0 difference coefficient captures listeners' sensitivity to direction of the pitch movement. Its being significantly positive for all groups means that the steeper the f_0 rise (including negative slopes) the longer the stimulus is judged by our subjects. Consequently, speakers judged rising stimuli as longer than the falling ones. The effect size differences between groups are not significant.

The absolute slope values (i.e., *absolute* dynamicity of stimulus, $|\Delta f_0|$ difference) significantly influenced Finnish and Mandarin speakers, but not the other two groups. The effect size was significantly smaller for EST than for MAN and FIN ($p < 0.01$).

The combined effect of Δf_0 and $|\Delta f_0|$ difference terms can be summarized as follows. For Estonians and Swedes, with negligible absolute dynamicity effect, the more positive (less negative) the slope, the greater durational judgment; in particular, the falls were judged shorter the steeper their negative slope. For the Finnish and Mandarin speakers, the absolute dynamicity effect magnifies the lengthening judgements for rising stimuli, but counteracts or even overturns the judged lengthening for falling tones. In effect, the falling tone would be judged equally long (FIN values) or even longer (MAN) than a level tone of the same physical duration.

Finally, positive values associated with the level difference mean that all groups were influenced by sound intensity in the expected direction (for SWE this effect was not statistically significant). The differences in effect sizes were not significant except EST–MAN pair ($p < 0.05$).

4. DISCUSSION AND CONCLUSIONS

Primarily, these results show that pitch level and intensity influences judgment of interval duration for speakers of several languages from different languages families in an expected way: the higher pitched and louder stimuli were judged as longer.

Although the spectral characteristics of our stimuli are considerably different from speech sounds, as mentioned above, they share several important features with phones of spoken language. Therefore, it is likely that the judgements reported in this paper provide relevant insights into how pitch level and duration interact in speech processing by humans.

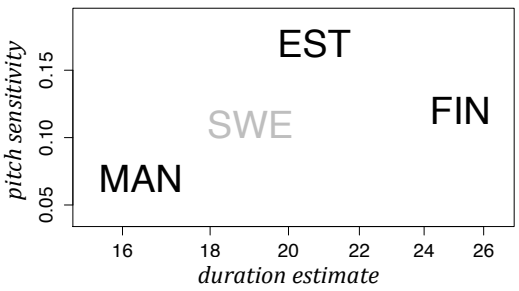

Figure 1: Distribution of languages based on estimates of duration and pitch sensitivity.

In fact, our results provide an intriguing support for such a claim. Fig. 1 shows the languages distributed based on duration and pitch level regression coefficients listed in Tab. 1. The closely related Finnic languages are in the top right corner (as the difference in pitch sensitivity was not significant for this pair, y-axis is plotted in log-scale). The greater precision in durational judgements may

be associated with the rich quantity systems in these languages. Moreover, the fact that both languages use pitch to co-signal phonological quantity might be related to greater influence of pitch level on duration estimates. Even within this cluster the significantly higher ability of duration discrimination ability by the Finns than Estonians might be linked to the differences in the quantity systems between these languages, with the Finnish system possibly relying more on single-phone duration compared to the Estonian one with robust compensatory effects.

Speakers of tonal Mandarin, which does not use duration phonologically, are less precise in durational judgments, and are to a smaller degree influenced by the interaction between duration and pitch. Swedish group, whose languages incorporates some aspects of both quantity and tonal contrast is placed in the middle of the figure (the grey color of SWE label due to the rather small dataset for this group).

To some extent, our results account well with the hyper-correction hypothesis discussed in Introduction. For example, the Mandarin data broadly agree with the perception results reported in [7] (except LH tones being judged as longer than HL by our subjects). For Estonian and Finnish, we can argue that long vowels are associated with lower f_0, due to the usually falling pitch pattern, and speakers of these languages hyper-correct their judgement of the lower sounds as being longer. Furthermore, the language specificity of the effect sizes shown here might be seen as an evidence for production based hyper-correction hypothesis claiming that speakers *judgments* are influenced by regularities in the language environment, and do not necessarily tell us much about low-level perceptual system.

The reported dynamicity effects are more difficult to interpret in this way. The speakers of Estonian, in which contrast between Q2–Q3 is associated with a falling f_0 contour (although not of the shape used for our stimuli) [17], indeed judged the falling tones as shorter, in accordance with the hyper-correction assumptions. But for the Finns, who in their language uniformly mark the distinction between two quantity levels by falling pitch in the long quantity, our data contradict the hyper-correction: they did not judge the stimuli with falling pitch as shorter. It is conceivable to interpret the Finnish pattern as a sign of an adaptation. A possible auditory bias to judge falling tones as shorter could impact the ability to make a correct quantity evaluation and, consequently, force the speakers to exaggerate the contrast by durational means, and is therefore suppressed. The reason why the Estonian judgements are more in line with hyper-correction hypothesis might be due to the more complex tonal marking of quantity in that language. Even the disagreement mentioned above between our results for Mandarin and those reported in [7] might be seen in this light: the assumed tendency of an auditory mechanism to judge falling tones – already short due to production constraints – as even shorter is to some extent offset by the strong absolute dynamicity effect, albeit not fully for our non-linguistic stimuli.

Furthermore, the existence of universal auditory biases is supported by the fact that, despite the typological differences, the signs of the significant coefficients for each individual effect (Tab. 1) are the same for all four languages. As brainstem level EEG measurements have shown, group level differences among languages in effect size are possible even at this processing stage (see e.g. [12, 5]); the auditory biases can be shaped by language environment leading to language specific modulation strengths.

It is therefore plausible that natural auditory biases serve as a basis for production patterns piggybacking on the properties of perception apparatus by hyper-correction and other mechanisms [28]. The resulting production regularities might then reinforce the biases neurally; plasticity of brainstem neurophysiology during language acquisition may play a role in this process. This in turn may lead to subsequent adjustments to production patterns, etc.

Different languages might take slightly different turns during their evolution; these random variations get reinforced through the process and may lead not only to known language differences but might also have an influence on the results of perception experiments such as reported here.

One way to contribute to verification of this hypothesis is to collect data from more languages and check whether (1) the main perceptual biases remain significant and in the expected direction, (2) the mutual relations revealed by perceptual / judgmental biases remain meaningfully linked to relative characteristics of languages as illustrated in Fig. 1. In addition, it is possible that using a different type of stimuli – more or less speech-like sounds embedded in more or less linguistically meaningful context – would heighten or lessen the differences between the listeners from different language background.

5. ACKNOWLEDGEMENTS

We thank Seila Pihanurmi, Katrin Leppik and Helen Türk for help with data collection. The work was in part funded by the Finnish Academy grant, Estonian Research Council grant IUT2-37 and Swedish Research Council project 2014-1072 *Andning i samtal*.

6. REFERENCES

[1] Brigner, W. L. 1988. Perceived duration as a function of pitch. *Perceptual and motor skills* 67(1), 301–302.

[2] Burghardt, H. 1973. Die subjektive Dauer schmalbandiger Schallen bei verschiedenen Frequenzlagen. *Acustica* 28, 278–284.

[3] Cooke, M. 1993. Modelling auditory processing and organisation.

[4] Cumming, R. 2011. The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics* 39(3), 375–387.

[5] Dawson, C., Aalto, D., Šimko, J., Putkinen, V., Tervaniemi, M., Vainio, M. 2014. Language-based plasticity in the auditory brainstem. The Neurosciences and Music-V: Cognitive Stimulation and Rehabilitation. Dijon, France.

[6] Gandour, J. 1977. On the interaction between tone and vowel length: Evidence from thai dialects. *Phonetica* 34(1), 54–65.

[7] Gussenhoven, C., Zhou, W. 2013. Revisiting pitch slope and height effects on perceived duration. *INTERSPEECH* 1365–1369.

[8] Järvikivi, J., Vainio, M., Aalto, D. 2010. Real-time correlates of phonological quantity reveal unitiy of tonal and non-tonal languages. *PloS one* 5(9), 619–639.

[9] Jeon, J. Y., Fricke, F. R. 1997. Duration of perceived and performed sounds. *Psychology of Music* 25(1), 70–83.

[10] Kinoshita, K., Behne, D. M., Arai, T. 2002. Duration and F0 as perceptual cues to Japanese vowel quantity. *Proccedings of the 7th Internationmal Conference on Spoken Language Processing*.

[11] Kong, Q.-M. 1987. Influence of tones upon vowel duration in Cantonese. *Language and Speech* 30(4), 387–399.

[12] Krishnan, A., Xu, Y., Gandour, J., Cariani, P. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. *Cogn. Brain Res.* 161–168.

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

[14] Lehnert-LeHouillier, H. 2007. The influence of dynamic f0 on the perception of vowel duration: Cross-linguistic evidence. *Proceedings of the 16th International Congress of Phonetic Sciences* 757–760.

[15] Lindblom, B. 1990. Explaining Phonetic Variation: A Sketch of the H&H Theory. In: Hardcastle, W. J., Marchal, A., (eds), *Speech Production and Speech Modelling*. Kluwer Academic Publishers 403–439.

[16] Lippus, P., Pajusalu, K., Allik, J. 2009. The tonal component of Estonian quantity in native and non-native perception. *Journal of Phonetics* 37(4), 388–396.

[17] Lippus, P., Pajusalu, K., Allik, J. 2011. The role of pitch cue in the perception of the Estonian long quantity. In: *Prosodic categories: Production, perception and comprehension*. Springer 231–242.

[18] Norman, J. 1988. *Chinese*. Cambridge University Press.

[19] Ohala, J. J. 1993. The phonetics of sound change. *Historical linguistics: Problems and perspectives* 237–278.

[20] Pisoni, D. B. 1976. Fundamental frequency and perceived vowel duration. *The Journal of the Acoustical Society of America* 59(S1), S39–S39.

[21] Rosen, S. M. 1977. The effect of fundamental frequency patterns on perceived duration. In: *Speech Transmission Laboratory—Quarterly Progress and Status Report* volume 18. Stockholm, Sweden: KTH 17–30.

[22] Suomi, K. 2005. Temporal conspiracies for a tonal end: Segmental durations and accentual f0 movement in a quantity language. *Journal of Phonetics* 33(3), 291–309.

[23] Takiguchi, I., Takeyasu, H., Giriko, M. 2010. Effects of a dynamic f0 on the perceived vowel duration in japanese. *Proceedings of the 5th International Conference on Speech Prosody*.

[24] Vainio, M., Järvikivi, J., Aalto, D., Suni, A. 2010. Phonetic tone signals phonological quantity and word structure. *The Journal of the Acoustical Society of America* 128, 1313.

[25] Vainio, M., Järvikivi, J., Aalto, D., Suni, A. 2010. Phonetic tone signals phonological quantity and word structure. *J. Acoust. Soc. Am.* 128, 1313–1321.

[26] Van Dommelen, W. 1995. Interactions of fundamental frequency contour and perceived duration in Norwegian. *Phonetica* 52(3), 180–187.

[27] Wang, W. S.-Y., Lehiste, I., Chuang, C.-K., Darnovsky, N. 1976. Perception of vowel duration. *The Journal of the Acoustical Society of America* 60(S1), S92–S92.

[28] Yu, A. C. 2010. Tonal effects on perceived vowel duration. *Laboratory Phonology* 10, 151–168.