The effect of harmonic number and pitch salience on the ability to understand speech-on-speech based on differences in fundamental frequency

SARA M. K. MADSEN[1,*], TORSTEN DAU[2] AND ANDREW J. OXENHAM[1]

[1] *Department of Psychology, University of Minnesota, 75 East River Parkway, Minneapolis, MN, 55455, USA*

[2] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Ørsteds Plads, Building 352, 2800 Lyngby, Denmark*

Differences in fundamental frequency (F0) between competing voices facilitate the ability to segregate a target voice from interferers, thereby enhancing speech intelligibility. Although lower-numbered harmonics produce greater pitch salience than higher-numbered harmonics, it remains unclear whether differences in harmonic ranks, and therefore pitch salience, affect the benefit of pitch differences. Earlier studies have not reported an effect of pitch salience, but have generally used only conditions where the difference in average F0 (ΔF0) between the two competing voices was large. It is possible that the effect of pitch salience is greater in more challenging conditions, in which the ΔF0 is relatively small. This study tested speech intelligibility in the presence of one speech masker for ΔF0s of 0, 2, and 4 semitones. The speech was presented in a broadband condition or was highpass or lowpass filtered to manipulate the pitch salience of the voicing. Results showed no interaction between filter type and ΔF0, suggesting little or no effect of harmonic rank or pitch salience in the ability to use F0 to segregate voices, even with smaller ΔF0s between competing voices. The results suggest some benefit of ΔF0 between competing voices, even in the absence of low-numbered spectrally resolved harmonics.

INTRODUCTION

Pitch differences between competing voices can enhance our ability to segregate target speech from a background of other speakers (Bird and Darwin, 1998; Brokx and Nooteboom, 1982). It is, for example, easier to understand a female speaker masked by a male speaker than by another female speaker. Therefore, it seems plausible that the ability to make use of pitch differences would improve with the strength (salience) of the pitch of the speech. However, the importance of pitch salience for understanding speech in a speech background is unclear.

Pitch salience has been shown to affect the ability to discriminate small differences in F0 between consecutive complex tones in studies that varied the pitch salience by

*Corresponding author: madse399@umn.edu

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. © The Authors. ISSN: 2596-5522.

varying the number of the lowest harmonic component (rank; low harmonic rank is associated with high pitch salience and vice versa) included in the stimuli (Bernstein and Oxenham, 2006; Hoekstra and Ritsma, 1977; Shackleton and Carlyon, 1994). More specifically, these studies found that thresholds are lowest for tones with low harmonic rank and increase with increasing lowest harmonic rank, often reaching a plateau when the lowest harmonic present exceeds the 10th. This is also the point at which no harmonics are thought to be spectrally resolved, although the link between spectral resolvability and harmonic number remains uncertain (Bernstein and Oxenham, 2003; Graves and Oxenham, 2019).

Psychoacoustic studies of sound segregation based on fundamental frequency (F0) differences have often been carried out with interleaved sequences of tones. One such study did not find a difference in amount of perceived segregation as a function of F0 difference between tones consisting only of high harmonic ranks and tones containing low harmonic ranks (Vliegen and Oxenham, 1999), whereas others did find a significant effect of harmonic rank on segregation (Grimault *et al.*, 2000; Madsen *et al.*, 2018). Moreover, one of the studies found a correlation between F0 difference limens (DLs) and performance in a stream segregation task (Madsen *et al.,* 2018), supporting the idea that perceptual salience of cues used for segregation is important for the ability to segregate sounds.

However, even if pitch salience is important for sound segregation, it is not necessarily important for speech-on-speech perception. One study explored the effect of harmonic rank on speech intelligibility by comparing conditions where the target and one speech masker had been either lowpass (LP) or highpass (HP) filtered to retain or remove resolved components, respectively (Oxenham and Simonson, 2009). Surprisingly, similar speech intelligibility and masking release were found in both conditions, suggesting no benefit of having resolved harmonic components in the speech. However, this study only tested conditions where the long-term average F0 difference between the voices (ΔF0) was four or eight semitones (ST), according to recent F0 estimates obtained with Praat (Boersma and Weenink, 2009). It may be that pitch salience is only relevant for more challenging conditions, i.e. for conditions with smaller values of ΔF0.

The aim of the present study was to determine whether there is an effect of harmonic rank on the ability of listeners to use differences in F0 between a target talker and an interfering speech masker to understand speech. Speech from a target speaker and a masker was either LP filtered (low harmonic rank condition) or HP filtered (high harmonic rank condition) and the masker was manipulated with Praat to obtain conditions where the long-term average F0 of the target and masker was separated by 0, 2, or 4 STs. F0DLs were measured for a subset of the participants to confirm that the filter cutoff frequencies used yielded conditions with and without spectrally resolved harmonics, yielding good and poor pitch discrimination, respectively.

METHODS

General methods

The experiments were conducted in a double-walled acoustically shielded booth. The stimuli were generated in MATLAB (The Mathworks, Natick, MA, USA) and were presented at a sampling rate of 48000 Hz via a Fireface UCX sound card (RME, Haimhausen Germany) and Sennheiser HD 650 headphones (Sennheiser, Wedemark, Germany). The experimental protocols were approved by the Scientific Ethical Committees of the Capital Region of Denmark (H-16036391).

Speech experiment

Eighteen native-Danish-speaking participants (9 female) between 20 and 28 years (mean = 23.67, SD = 2.45) were tested. All participants had audiometric thresholds at octave frequencies between 250 and 8000 Hz no greater than 20 dB HL.

Speech intelligibility was tested for sentences masked by one speech masker for conditions where the speech material was either HP filtered, LP filtered, or unfiltered (broadband). The masker speech was manipulated in Praat to generate conditions where the difference in average long-term F0 of the target and masker (ΔF0) varied. The target consisted of sentences from the CLUE speech corpus (Nielsen and Dau, 2009). These are short contextual sentences similar to the HINT sentences (Nilsson *et al.*, 1994) that had a duration between 1.23 and 1.86 s. Speech from recordings of conversations (Sørensen *et al.*, 2018) was used as maskers. The recordings from two speakers were concatenated separately and all gaps exceeding 100 ms, non-Danish words, loud exclamations, and other sounds such as laughter were removed. The maskers for the main test were generated from the remaining speech material from one of the speakers that had a duration of 222.3 s and was divided into 180 overlapping segments of 2.47 s. Similarly, the maskers for the training (30 blocks of 2.47 s) were made from the remaining speech material from the other speaker. The maskers started 500 ms before the target and ended at least 100 ms after the target and were gated with 50 ms raised-cosine onset and offset ramps. One CLUE sentence was presented in quiet immediately before each trial to guide the participant towards the target voice. The guide sentence was always the same and was never the same as the target sentence. All speakers were male. The long-term average F0 was approximately 110 Hz for the target, 139 Hz for the masker used for testing, and 148 Hz for the masker used for training. The F0s of the maskers were manipulated in Praat to obtain differences between the long-term average F0 of the target and masker (ΔF0) of 1, 3, and 5 STs for the training and 0, 2, and 4 STs for the main test. The average long-term F0 of the masker was always the same as or higher than that of the target. The speech maskers were filtered to have the same long-term spectrum as the CLUE sentences. For the HP- and LP-filtered conditions, the target and masker were filtered with a 4th-order Butterworth filter after being combined. The guide sentences were filtered with the same filter. The conditions with ΔF0 of 4 STs were used as a reference since it is the only ΔF0 tested both here and in the study by Oxenham and Simonson (2009).

Cutoff frequencies of 800 Hz and 1500 Hz were chosen for the LP- and HP-filtered conditions, respectively, since pilot experiments indicated that they would yield similar performance. A target-to-masker ratio (TMR) of 0 dB was used for the filtered conditions and a TMR of -15 dB was used for the broadband conditions to obtain similar performance in the filtered and broadband conditions for ΔF0 = 4 STs.

A Gaussian noise with the same long-term spectrum as the CLUE sentences (before filtering) was filtered with a 4th-order Butterworth filter and added to the filtered speech stimuli. For the LP-filtered condition, the noise was HP filtered with a cutoff frequency of 800 Hz and for the HP-filtered condition, the noise was LP filtered with a cutoff frequency of 1500 Hz. The level of the noise before filtering was 12 dB lower than the unfiltered target speech. The target and maskers combined were presented at an overall sound pressure level (SPL) of 70 dB.

In the main test, each of the nine conditions (three filter conditions and three ΔF0s) was tested with two lists each containing 10 sentences. The order of the conditions was randomized within each of two consecutive blocks, both containing all of the nine conditions. The training consisted of three runs presented in the following order: 1) Broadband with ΔF0 of 5 STs presented at a TMR of -12 dB; 2) HP filtered with ΔF0 of 3 STs, presented at a TMR of 3 dB; 3) LP filtered with ΔF0 of 1 ST, presented at a TMR of 0 dB.

The participants were instructed to listen for the voice of the guide sentence and were asked to type what they heard that voice said, after each trial. The speech scores were transformed into rationalized arcsine units (RAU) before statistical analysis.

F0 discrimination limens

F0 discrimination limens (F0DLs) were measured with a two-interval, three-down, one-up adaptive procedure where each interval contained four 200-ms tones, presented immediately after each other as in earlier studies (Madsen *et al.*, 2019; Madsen *et al.*, 2017). In the reference interval, all tones had the same F0, the reference F0, that was roved over two semitones and centred on 131 Hz (corresponding to one standard deviation above the long-term average F0 of the target speech). In the target interval, the F0 of the first and third tone was higher and the F0 of the second and fourth tone was lower than the reference F0. The difference in F0 between the high and low tones was varied adaptively, while the F0s of the tones remained geometrically centered on the reference F0. All tones were gated with raised-cosine ramps of 20 ms and the two intervals were separated by a 400 ms pause. The participants were asked to indicate which interval contained the changes in pitch. Feedback was provided after each trial.

The harmonic components were added in either sine or random phase. For the latter, the phase was for each component chosen randomly and independently from a uniform distribution from 0 to 2π. As in the speech experiment, the tones were either LP or HP filtered with a fourth-order Butterworth filter using cutoff frequencies of

800 Hz and 1500 Hz, respectively. Moreover, as in the speech experiment, a HP-filtered Gaussian noise with cutoff frequency of 800 Hz was added in the LP filtered condition and a LP-filtered Gaussian noise with cutoff frequency of 1500 Hz was added in the LP filtered condition. The average overall level of the tones was 70 dB SPL but the level was roved independently for each tone over a uniform range of 6 dB. The noise was presented at a level 12 dB below the nominal level of the tones before LP or HP filtering.

For each run, the thresholds were calculated as the geometric mean across the last six reversals. The experiment contained three blocks with one run for each condition and the order of the conditions were randomized within a block. The first run was used for training and the final thresholds were defined as the geometrical mean across the two last runs.

RESULTS

Speech intelligibility was measured as the proportion of words reported correctly in each condition. All deviations except obvious misspellings or homophones were considered incorrect. Additional words and differences in word order were not penalized.

The scores for the reference conditions (ΔF0 = 4 STs) and broadband conditions are shown in the left and middle panel of Fig. 1, respectively. It can be seen that the mean scores are very similar for the three reference conditions, and a repeated-measures ANOVA with filter type as the within-subjects factor found no significant effect of filter condition [F(2,34) = 0.47, p = 0.63]. This finding confirmed that the cutoff frequencies chosen for the HP- and LP-filtered conditions and the TMRs chosen for the filtered and broadband conditions yielded similar performance in the three reference conditions. The speech scores for the broadband conditions were analyzed separately since the filtered conditions were measured at a higher TMR than the broadband conditions. For the broadband condition, there was a tendency for the scores to increase slightly with increasing ΔF0 even though the scores for ΔF0 = 0 and ΔF0 = 2 were very similar to each other. Analysis of the speech scores with ΔF0 as the within-subject factor showed a small but significant effect of ΔF0 [F(2,34) = 3.35, p = 0.047]. Moreover, Bonferroni-corrected post-hoc tests showed a significant difference between the conditions with ΔF0 of 0 and 4 semitones [t(34) = -2.55, p = 0.046] but not between either of the other pairs of conditions. The right panel of Fig. 1 shows individual and mean speech scores for the LP- and HP-filtered conditions for ΔF0s of 0, 2, and 4 STs. As expected, scores generally increased with increasing ΔF0. Furthermore, the scores were generally higher for the HP than for the corresponding LP conditions especially for ΔF0 = 0 STs and for ΔF0 = 2STs. Analysis with ΔF0 and filter condition as within-subject factors show a significant effect of both ΔF0 [F(2, 34) = 24.37, p < 0.0001] and filter condition [F(1,17) = 9.51, p = 0.0067] but no interaction between ΔF0 and filter condition [F(2, 34) = 1.25, p = 0.3], indicating that low and high harmonics both facilitate improvements in performance with F0 differences at similar rates in the presence of competing voices.

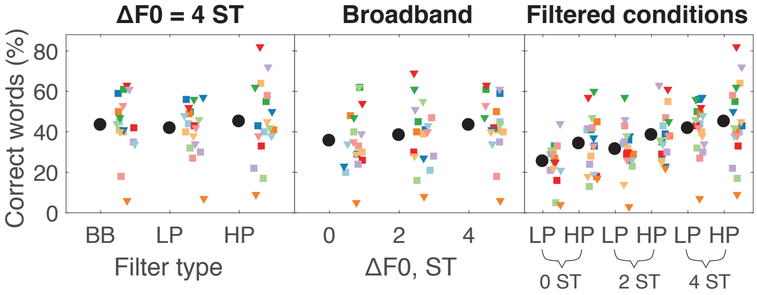

Fig. 1: Speech scores for the reference conditions (left panel), the broadband condition (middle panel), and the LP- and HP-filtered conditions (right panel). Larger circles represent the mean across participants and the smaller symbols show the individual scores.

F0 discrimination

F0 discrimination thresholds are shown in Fig. 2. As expected, F0DLs were higher in the HP conditions than in the LP conditions, and phase affected F0DLs in the HP, but not in the LP, conditions. There were significant effects of both filter type [F(1, 5) = 40.51, p = 0.00011] and phase [F(1,5) = 11.99, p = 0.018] and a significant interaction between phase and filter condition [F(1,5) = 7.60, p = 0.040]. The results are consistent with expectations based on high-ranked unresolved harmonics being present when the stimuli were HP filtered with a cutoff frequency of 1500 Hz, as in the speech experiment. The results confirm that the pitch of the speech was less salient under HP conditions than under LP conditions.

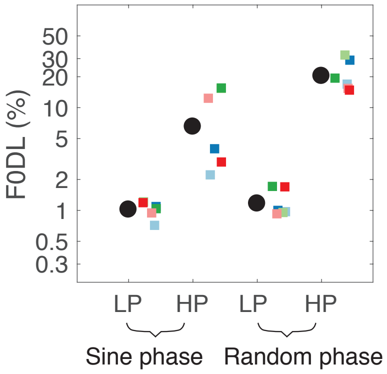

Fig. 2: F0 discrimination thresholds for complex tones with components added in sine or random phase and then LP or HP filtered at 800 Hz or 1500 Hz, respectively.

DISCUSSION

In the speech experiment, there were significant effects of ΔF0 and filter type but no significant interaction between ΔF0 and filter condition. The improvement in speech scores with increasing ΔF0 is consistent with results from earlier studies (Bird and Darwin, 1998; Brokx and Nooteboom, 1982; Madsen *et al.,* 2017). The lack of a main effect of filter type when considering the reference conditions confirms that we were successful at selecting filter cutoff frequencies that produced roughly equal performance in the LP and HP conditions when ΔF0 = 4 ST. Due to this equalization of performance, it is more relevant to compare the slopes of the scores for the HP-filtered conditions relative to the LP-filtered conditions as a function of ΔF0 instead of the absolute scores. The seemingly steeper slope for the LP condition compared to the HP condition suggests that pitch is more important for the LP than for the HP condition. This support the idea that the more salient pitch cues in the LP condition lead to better separation of the voices and therefore better speech intelligibility. However, the lack of a significant interaction between filter condition and ΔF0 indicates that this effect is not robust. The lack of interaction is consistent with the results from Oxenham and Simonson (2009), which showed similar performance for a HP- and a LP-filtered conditions for ΔF0s of 4 STs and 8 STs. This may be explained by the different forms of speech information conveyed in the low and high spectral region, respectively, or by the difference strength of masker modulation in the two spectral regions as proposed by Oxenham and Simonson (2009). Another possible explanation is that, despite testing the smallest possible long-term average ΔF0 of 0 STs, the momentary differences in ΔF0 might have been too large for differences in pitch salience to affect speech intelligibility. This would suggest that pitch salience would not be an issue for understanding speech-on-speech in real-life situations.

In summary, this study tested speech intelligibility in a background of a speech masker and found a small effect of ΔF0 but a similar relation between performance in LP- and HP-filtered conditions for different ΔF0s. This suggest that the difference in pitch salience between low-numbered and high-numbered harmonics is not a determining factor for the ability to use F0 differences between competing talkers to better understand speech.

ACKNOWLEDGEMENTS

Work supported by the Demant Foundation and NIH grant R01 DC005216.

REFERENCES

Bernstein, J.G.W., & Oxenham, A.J. (**2003**). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?" J. Acoust. Soc. Am., **113**(6), 3323–3334. doi: 10.1121/1.1572146

Bernstein, J.G.W., & Oxenham, A.J. (**2006**). "The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss," *J.* Acoust. Soc. Am., **120**(6), 3929–3945. doi: 10.1121/1.2372452

Bird, J., & Darwin, C.J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," *Psychophysical and Physiological Advances in Hearing*, 263–269.

Boersma, P., & Weenink, D. (**2009**). Praat: Doing phonetics by computer (Version 5.1.3.1).

Brokx, J.P.L., & Nooteboom, S.G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon., **10**, 23–36.

Graves, J.E., & Oxenham, A.J. (**2019**). "Pitch discrimination with mixtures of three concurrent harmonic complexes," J. Acoust. Soc. Am., **145**(4), 2072-2083.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., & Collet, L. (**2000**). "Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency," J. Acoust. Soc. Am., **108**(1), 263–271. doi: 10.1121/1.429462

Hoekstra, A., & Ritsma, R.J. (**1977**). "Perceptive hearing loss and frequency selectivity," *Psychophysics and Physiology of Hearing*, 263–271.

Madsen, S.M.K., Dau, T., & Moore, B.C.J. (**2018**). "Effect of harmonic rank on sequential sound segregation," Hear. Res., **367**, 161–168. doi: 10.1016/j.heares.2018.06.002

Madsen, S.M.K., Marschall, M., Dau, T., & Oxenham, A.J. (**2019**). "Speech perception is similar for musicians and non-musicians across a wide range of conditions," Sci. Rep., **9**(1), 1-10. doi: 10.1038/s41598-019-46728-1

Madsen, S.M.K., Whiteford, K.L., & Oxenham, A.J. (**2017**). "Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds," Sci. Rep., **7**, 1-9. doi: 10.1038/s41598-017-12937-9

Nielsen, J.B., & Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol., **48**(10), 729–741. doi: 10.1080/14992020903019312

Nilsson, M., Soli, S.D., & Sullivan, J.A. (**1994**). "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., **95**, 1085–1099. doi: 10.1121/1.408469

Oxenham, A.J., & Simonson, A.M. (**2009**). "Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am., **125**(1), 457–468. doi: 10.1121/1.3021299

Shackleton, T.M., & Carlyon, R.P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," J. Acoust. Soc. Am., **95**(6), 3529–3540. doi: 10.1121/1.409970

Sørensen, A.J., Fereczkowski, M., & MacDonald, E.N. (**2018**). "Task dialog by native-Danish talkers in Danish and English in both quiet and noise," Zenodo. doi: 10.5281/zenodo.1204951

Vliegen, J., & Oxenham, A.J. (**1999**). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am., **105**, 339–346. doi: 10.1121/1.424503