Assessing the impact of fundamental frequency on speech intelligibility in competing-talker scenarios

PAOLO A. MESIANO[1,*], JOHANNES ZAAR[1], LARS BRAMSLØW[2], NIELS H. PONTOPPIDAN[2], AND TORSTEN DAU[1]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Augmented Hearing, Eriksholm Research Centre, 3070 Snekkersten, Denmark*

When only monaural cues are available in competing-talker scenarios, normal-hearing (NH) listeners are able to identify and understand the target speech while hearing-impaired listeners often experience difficulties. A good understanding of the role of monaural cues in speech segregation is therefore essential for developing hearing-aid compensation strategies. Earlier studies with NH listeners showed that differences in fundamental frequency (ΔF_0) between the target talker and one interfering talker can facilitate the segregation of the speech signals. However, most of these studies used speech materials that bear little resemblance with everyday speech. Furthermore, the F_0 was either defined by talker sex or measured as a talker-specific average, thus ignoring the significant F_0 variability across sentences. The present study instead used everyday-speech type sentences from the Danish Hearing in Noise Test (HINT) and employed a more accurate method for assessing the impact of F_0 on intelligibility for NH listeners. Compared to previous studies, the overall effect of ΔF_0 was found to be smaller and it was hypothesised that the previously employed speech materials might have enhanced the effect of ΔF_0 beyond its real-life importance.

INTRODUCTION

When several talkers speak simultaneously, it can be challenging to identify and understand one specific target-speech signal. In such situations, usually referred to as competing-talker scenarios, the healthy auditory system shows exceptional abilities of segregating the target speech from the interfering speech by making use of several auditory cues. For example, binaural cues are known to be beneficial for separating signals arriving from different directions. However, in situations where binaural cues are not present or unreliable, the auditory system must rely on monaural cues only. In fact, it has been demonstrated that, in such conditions, normal-hearing (NH) listeners are still able to identify and understand the target speech, whereas hearing-impaired (HI) listeners typically experience substantial difficulties (Bramsløw *et al.*, 2015). The fundamental frequency (F_0) is a monaural auditory cue with a strong impact on competing-talker scenarios: when the target speech signal and a competing speech

*Corresponding author: pamesi@dtu.dk

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. ISSN: 2596-5522.

signal differ in F_0, the perception of the target is typically facilitated (e.g., Brokx and Nooteboom, 1982, Darwin *et al.*, 2003, Summers and Leek, 1998, Assmann and Summerfield, 1990, Assmann, 1999). Several approaches have been used to investigate the role of F_0 differences between two competing talkers. However, the employed experimental scenarios have typically been different from everyday-life listening situations.

For example, in Brokx and Nooteboom (1982) and Darwin *et al.* (2003), the F_0 of the competing sentences was assessed as a talker-specific average, thus ignoring the considerable F_0 variability across sentences spoken by a given talker. Brokx and Nooteboom (1982) generated the F_0-separated condition by pairing target speech from a male talker with interfering speech from the same talker who was asked to imitate the (higher) pitch of a female voice. The resulting interfering signal had a higher F_0 on average, but the F_0 contour (i.e., F_0 as a function of time) showed large variations, often crossing the F_0 contour of the target. To obtain a more controlled F_0 difference between competing voices, Summers and Leek (1998) employed monotonized F_0 contours (constant F_0 over time). The advantage of this approach is a perfectly controlled F_0 difference for each pair of sentences, but the monotonized voice may sound unnatural to the listener. In Assmann (1999), pairs of sentences with naturally-varying F_0 contours were used and F_0 separations were generated as the difference between the across-time average F_0 of each sentence, rather than based on the average of the talker, obtaining a more accurate control of F_0 separation. However, Assmann (1999) aligned the competing sentences at their offsets, potentially introducing a strong cue due to the onset differences.

Furthermore, the speech corpora have typically been chosen to maximize the influence of the F_0 separation. Darwin *et al.* (2003) used competing sentences from the coordinate-response measure (CRM), a speech corpus comprised of time-aligned closed-set sentences with a pre-defined structure and two scoring keywords per sentence. They observed a significant improvement in speech intelligibility when two competing sentences differed in F_0 by more than two semitones and reported performance improvements of more than 20% for a 9-semitone separation. However, these strong benefits induced by F_0 differences might be exaggerated in relation to real-life speech, due to the high degree of word alignment in the CRM corpus.

The present study aimed to overcome the above mentioned limitations by (i) employing a more accurate method for measuring and generating the F_0 separation between competing sentences, taking into account the variability across sentences for each talker, and by (ii) using a more realistic and less constrained speech corpus as compared to the mentioned reference studies.

METHOD

The Danish Hearing in Noise Test (HINT) (Nielsen and Dau, 2011) was used to generate the experimental stimuli. The HINT speech corpus consists of 200 open-set five-word everyday-type natural sentences split into ten phonetically-balanced

lists. Recordings of the speech material from twelve different talkers (six males and six females) were used. First, F_0 contours were extracted with the software Praat (Boersma, 1993): the instantaneous F_0 was estimated in 10-ms time steps, within a frequency range between 30 and 550 Hz. Then, the across-time median and standard deviation of the F_0 contour for each sentence was computed. The median F_0 of a talker can vary significantly across the speech corpus, by up to four semitones. Therefore, the F_0 separation between paired sentences was measured as the difference between their median F_0 values, regardless of the average F_0 of the specific talker.

The stimuli were generated by pairing sentences spoken by the same talker, taken from different lists, presented simultaneously and aligned at their onsets. The sentences were processed in Praat to obtain a difference in median F_0 (ΔF_0) of 0, 3, 6 or 12 semitones. Each value of the F_0 contour was multiplied by a scaling factor $s > 0$. This approach preserved the natural increase in frequency range observed for increasing median F_0. To avoid shifting sentences to extreme and unnatural F_0 values, the desired ΔF_0 difference was split across the two sentences in each pair, as indicated in Table 1, where the F_0 shifts in semitones refer to the separation from the talker's average F_0. In each sentence pair, the larger F_0 shift was applied towards higher frequencies ($s > 1$) if the average F_0 of the talker was above the average F_0 of the entire speech corpus, and towards lower frequencies ($s < 1$) otherwise. Figure 1 shows an example of a sentence pair with different ΔF_0s, indicated in the different panels.

ΔF_0	Talker F_0 > overall F_0		Talker F_0 < overall F_0	
	F_0 shift (s>1)	F_0 shift (s<1)	F_0 shift (s>1)	F_0 shift (s<1)
0	0	0	0	0
3	1	-2	2	-1
6	2	-4	4	-2
12	4	-8	8	-4

Table 1: F_0 shifts applied to each sentence in a pair to obtain the desired ΔF_0.

The stimuli were presented using the competing-voices test (CVT) framework developed by Bramsløw *et al.* (2019), where the listeners were provided with the first word of the target sentence on a screen prior to the stimulus playback (text pre-cue). After playback, the listeners were asked to repeat as many words as possible from the target sentence. The target sentence was presented at an average sound pressure level (SPL) of 65 dB with level roving of $\pm$5 dB. Target-to-masker ratios (TMRs) of -12, -8, -4, 0, 4 dB were combined with the four ΔF_0 values to produce a total of 20 experimental conditions. Each condition was tested using 20 sentence pairs. The performance in the experiment was measured as the proportion of correctly repeated words in the target sentence, averaged over the 20 sentence pairs presented in each experimental condition.

To avoid any effect of presentation order or sentence repetition in the group results, the test conditions (ΔF_0 and TMR) were balanced across listeners using a Latin square

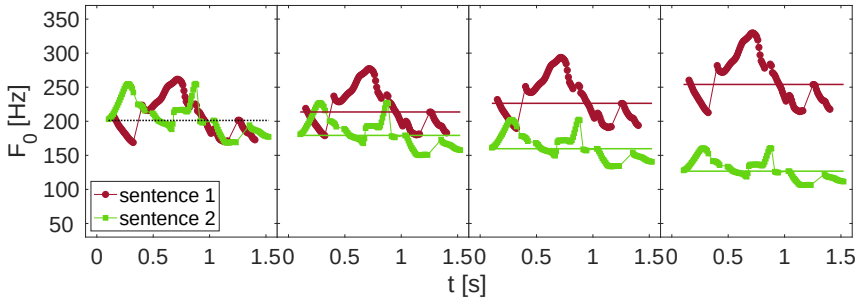

Fig. 1: Example of F_0 contour processing for a pair of HINT sentences. Panels from left to right show ΔF_0 of 0, 3, 6 and 12 semitones. The median F_0 of each sentence is displayed as a straight line, except for $\Delta F_0 = 0$ semitones (left panel) where the median F_0 of both sentences coincide with the talker's median F_0, shown as a dashed black line.

design while sentence-list and talker were randomized across conditions. In each pair, the target sentence was assigned randomly to either the sentence with higher or lower median F_0. Given the limited number of sentences in the HINT speech corpus, sentences have been repeated during the experiment, either as target or masker. However, the process of generating sentence pairs was randomized in a way that a sentence was presented a second time as late as possible and never within a given experimental condition.

The stimuli were free-field equalized and presented diotically over Sennheiser HDA200 headphones to the listeners seated in a sound-proof booth. Fifteen young native speakers of Danish with NH (pure-tone thresholds below 20 dB hearing level between 125 Hz and 8 kHz) were tested in the experiment.

RESULTS

The left panel of Figure 2 shows average group results, with the proportion of correct words displayed as a function of TMR. Each function represents results obtained for a particular ΔF_0 condition. Overall, speech intelligibility was found to improve with increasing TMR for all ΔF_0 values. The strongest effect of ΔF_0 was observed for intermediate TMR values (at -8 and -4 dB), with a maximum effect at TMR=-8 dB where the performance improved with increasing ΔF_0. At this TMR, the proportion of correct words increased by 15% from $\Delta F_0 = 0$ semitones to $\Delta F_0 = 12$ semitones. At the limits of the TMR range tested, only a minor effect of ΔF_0 was measured.

For comparison, the data from Darwin *et al.* (2003) are shown in the right panel of Figure 2. The range of TMR values in common between their experiment and the present study is indicated by the grey area in the two panels. Darwin *et al.* (2003) observed an overall stronger effect of ΔF_0 between target and interfering sentences, with the largest speech intelligibility improvement of about 30% at TMR = -3 dB for a separation of 12 semitones. In their study, the F_0 separation was effective for TMRs

ranging from -6 dB to 3 dB, decreasing for increasing TMR, and becoming negligible at the higher applied TMRs where a ceiling effect was observed.

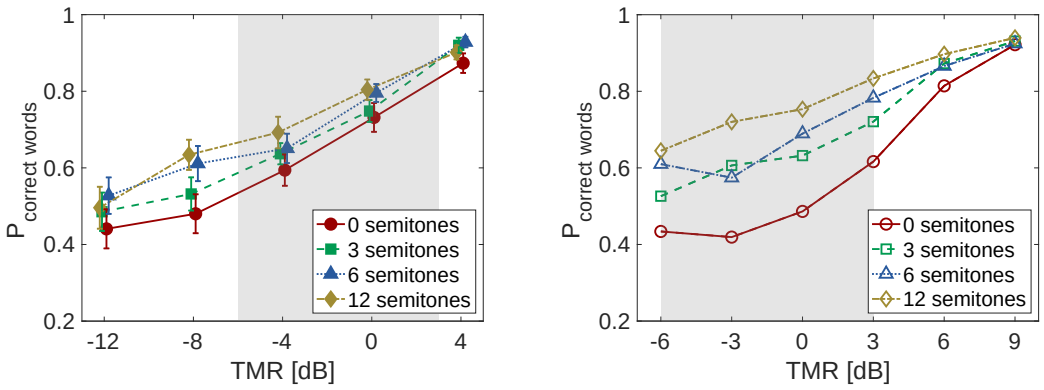

Fig. 2: Left panel: Group average proportion of correct words as a function of TMR. Different ΔF_0 conditions are shown as different curves. Error bars represent standard errors across listeners. Right panel: Data from Darwin *et al.* (2003) for comparison. The grey areas in the two panels indicate the group of TMR values that was common in the two studies.

A mixed-model analysis of variance (ANOVA) was conducted on rau-transformed data, including listener, ΔF_0 and TMR as main factors. The listener was treated as a random factor while ΔF_0 and TMR as fixed factors. To analyse the interactions between the main experimental factors, two-way interactions were also included. The results of the ANOVA are reported in Table 2. All main factors were significant ($p < 0.001$), indicating that performance differs across listeners and that both ΔF_0 and TMR affect performance significantly. However, no significant interactions between factors were observed.

Factor	**Fixed/Random**	**df**	**F**	**p**
Listener	Random	14	11.17	<0.001
ΔF_0	Fixed	3	7.85	<0.001
TMR	Fixed	4	118.24	<0.001
Listener*ΔF_0	Random	42	1.21	0.201
Listener*TMR	Random	56	1.4	0.055
ΔF_0*TMR	Fixed	12	0.77	0.685
Error		167		

Table 2: Results of mixed-model analysis of variance performed on RAU-transformed proportion of correctly repeated words.

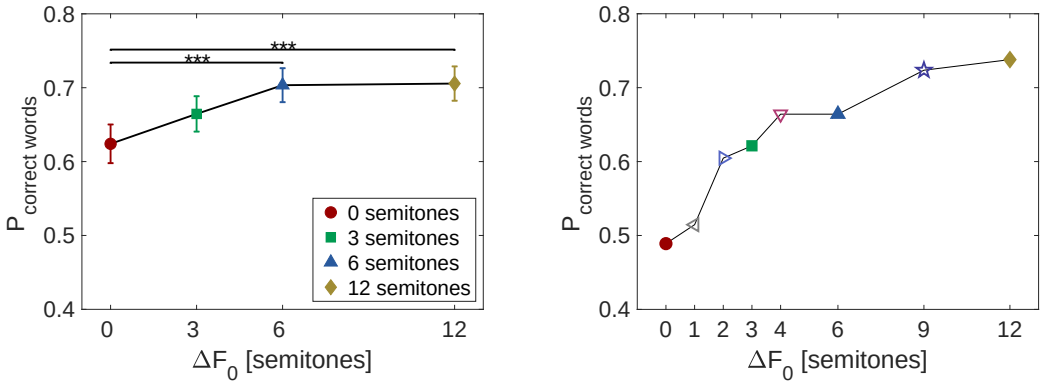

Fig. 3: Left panel: Word-recognition performance as a function of ΔF_0, averaged across TMRs. Error bars represent 95% confidence intervals. Right panel: Data from Darwin *et al.* (2003) averaged across their four lowest TMRs. The filled symbols represent ΔF_0 values that were in common with those of the current study.

Figure 3 shows the word recognition performance as a function of ΔF_0, averaged across TMRs. The results of the present study are shown in the left panel and those by Darwin *et al.* (2003) are displayed in the right panel. A post-hoc pairwise comparison analysis of the data from the present study showed that the difference between $\Delta F_0 = 0$ semitones and ΔF_0 = 6 or 12 semitones was statistically significant at the $p < 0.001$ level. The effect of ΔF_0 increased by 8% for ΔF_0 = 6 semitones and saturated above it, such that an increase in F_0 separation above 6 semitones did not provide any additional increase of speech intelligibility. In contrast, the data from Darwin *et al.* (2003), averaged across their four lowest TMRs, showed that a two-semitone separation was sufficient to obtain a 12% increase of speech intelligibility relative to the zero-semitone separation condition.

DISCUSSION

Overall, the results from the present study demonstrate that the benefit of substantial F_0 separations between competing sentences is relatively small in NH listeners when using a realistic speech corpus, in particular at conversational (positive) TMRs. The strongest effect of ΔF_0 was found at the negative TMRs and no effect of ΔF_0 was observed for the extreme values of TMR (-12 dB and 4 dB). It is possible that the task was too difficult at TMR = -12 dB, therefore making the ΔF_0 information ineffective, whereas it was too easy at TMR = 4 dB, making the ΔF_0 cue superfluous.

A strictly monotonic increase of speech intelligibility with increasing TMR was found for all ΔF_0s. This is in contrast to previous findings where results showed a non-monotonic relation with a local minimum at about 0 dB TMR (Brungart, 2001). This non-monotonic behavior was attributed to the low target level at negative TMRs that might have facilitated the segregation cue.

The high degree of sentence synchrony in the CRM speech corpus might have led to large amounts of energetic masking and thus to an overall more challenging task compared to a similar experiment that employed the HINT sentences. This can be noticed by comparing performances for $\Delta F_0 = 0$ semitones between the current and the reference study (left and right panels in Figure 2, respectively). Performances observed by Darwin *et al.* (2003) are lower than those obtained with the HINT speech corpus at the same TMR, meaning that to obtain a given speech intelligibility, more beneficial TMRs were needed in the reference study. A high degree of energetic masking may have emphasised the effect of F_0 separation beyond its real-life importance, resulting in an overestimation of the effects of ΔF_0.

Further work is required to prove this hypothesis. Furthermore, additional research is needed to investigate which other cues at the level of the F_0 contribute to speech separation besides the ΔF_0 and how these cues are affected by hearing impairment. The results may then be relevant for the development of hearing-aid processing strategies targeted to restore speech intelligibility in competing-talker scenarios.

REFERENCES

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am., **88**(2), 680-697.

Assmann, P. F. (**1999**). "Fundamental frequency and the intelligibility of competing voices," in Proc. International Congress of Phonetic Sciences, 179-182.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in Proceedings of the Institute of Phonetic Sciences, **17**(1193), 97-110.

Bramsløw, L., Vatti, M., Hietkamp, R.K. and Pontoppidan, N. (**2015**). "Binaural speech recognition for normal-hearing and hearing-impaired listeners in a competing voice test," Proc. Speech in Noise, Copenhagen, Denmark.

Bramsløw, L., Vatti, M., Rossing, R., Naithani, G., and Henrik Pontoppidan, N. (**2019**). "A Competing Voices Test for Hearing-Impaired Listeners Applied to Spatial Separation and Ideal Time-Frequency Masks," Trends Hear., **23**, 1-12.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon., **1**(1), 23-36.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am., **109**(3), 1101-1109.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am., **114**(5), 2913-2922.

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**(3), 202-208.

Summers, V., and Leek, M. R. (**1998**). "F0 processing and the seperation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res., **41**(6), 1294-1306.