Speech modifications in interactive speech: Effects of age, sex and noise type.

Outi Tuomainen[1,2], Linda Taschenberger[1], Stuart Rosen[1], Valerie Hazan[1]

[1]University College London, Speech Hearing and Phonetic Sciences, 2 Wakefield Street, London WC1N 1PF, United Kingdom

[2] University of Potsdam, Department of Linguistics, Haus 14, Karl-Liebknecht-Straße 24-25

14476 Potsdam; https://orcid.org/0000-0002-8654-2446

Corresponding author email: tuomainen@uni-potsdam.de

Keywords: Speech production, Speech adaptations, Vocal effort, Communication effort, Lombard speech, Lifespan

Accepted for publication for the special issue on "Voice modulation: from origin and mechanism to social impact", Philos Trans R Soc Lond B Biol Sci., 3/6/2021.

doi: 10.1098/rstb.2020.0398

Summary

When attempting to maintain conversations in noisy communicative settings, talkers typically modify their speech to make themselves understood by the listener. In this study, we investigated the impact of background interference type and talker age on speech adaptations, vocal effort and communicative success. We measured speech acoustics (articulation rate, mid-frequency energy, fundamental frequency), vocal effort (correlation between mid-frequency energy and fundamental frequency) and task completion time in 114 participants aged 8-80 years carrying out an interactive problem-solving task in good and noisy listening conditions (quiet, non-speech noise, background speech). We found greater changes in fundamental frequency and mid-frequency energy in non-speech noise than in background speech, and similar reductions in articulation rate in both. However, older participants (50+ years) increased vocal effort in both background interference types whereas younger children (<13 years) increased vocal effort only in background speech. The presence of background interference did not lead to longer task completion times. These results suggest that when the background interference involve a higher cognitive load, as in the case of other speech of other talkers, children and older talkers need to exert more vocal effort to ensure successful communication. We discuss these findings within the communication effort framework.

Introduction

When conversing in less than ideal or "challenging" conditions, such as in background noise or with someone with a hearing loss, talkers continuously monitor the success of the communication. In a case of communication breakdown, they modify their speech in an attempt to make themselves more intelligible to the listener. These modifications include a range of acoustic-phonetic (e.g., slower, more intense and hyper-articulated speech) and linguistic adaptations (e.g., higher-frequency words, shorter and simpler sentences) often broadly referred to as "clear speech" (for reviews see [1,2]). It has been shown that these speech modifications are modulated by complex interactions between various talker-related (e.g., age, regional accent), listener-related (e.g., hearing acuity) and environment-related factors (e.g., room acoustics, background noise; see [1] for a review). In this study, we focus on a few of these factors, namely how speech modifications in challenging conditions vary as a function of **type of background noise** and **talker age**.

Interactive speech communication in background noise is governed by a complex array of intelligibility-enhancing dynamic vocal changes (known as Lombard speech) in response to both changes in the perception of one´s own voice and the level of understanding by the listener ([3,4]). Lombard speech is characterised by acoustic-phonetic changes (e.g., intensity, duration, and fundamental frequency) that are associated with increases in vocal effort to overcome the effect of interfering noise, and the magnitude of adjustment in vocal effort is largely dependent on the changes in the level of background noise. Lombard speech is partially involuntary and reflexive: when the talker is unable to hear their own voice, they automatically increase their vocal effort. However, these speech modifications are also governed by communicative intent, with greater voice modifications observed when talkers are addressing their speech to someone as opposed to, for example, reading sentences ([3,4]). Although speech adaptations due to background noise are well documented in the literature, there is no clear consensus over the definition and operationalisation of vocal effort: adjustment of vocal effort is not limited to situations where there is background noise present and, therefore, it can be broadly defined as any exertion associated with speech production (see [5] for a discussion of terminology).

Another useful strategy to maintain conversation in a noisy communicative setting is to allocate more cognitive resources (attention, working memory) to what has been said by the conversation partner, that is, to increase listening effort, as accounted for by the *Framework for Understanding Effortful Listening* (FUEL; [6]). However, as with vocal effort, characterising and assessing listening effort has proven to be elusive ([6, 7, 8]; also see [9] for a review). More importantly, neither the framework for vocal effort (here: Lombard speech) nor for listening effort (FUEL) capture well the interactive nature of communicative situations that involve both speaking (planning and executing motor commands) and listening simultaneously (attending to yourself and the conversational partner). This dynamic production-perception behaviour in real-world situations is better captured by the term *communication effort* ([10]). In this framework, the main goal in interaction is to maintain mutual understanding while minimising vocal and listening effort to avoid fatigue (see also the *H&H model* [11]). Challenging acoustic conditions drive speech modifications, such as Lombard speech, that enhance the intelligibility of the speech for both the conversational partners and the talkers themselves. However, in order to maintain the appropriate level of vocal and listening effort required for successful communication, the dyads involved in the conversation need to constantly monitor each other´s feedback as to whether they were understood or not, which in turn then determines the expenditure of effort. This interactive production-perception loop is influenced by factors both external (e.g., the acoustic environment) and internal (sensory and cognitive abilities of the conversational partners).

Not all noises are equal, however, and different background noise *types* can have a differential impact on interactive speech communication. Background noise can physically interfere with the speech signal by direct interactions of energy in the target and masker (so-called *energetic masking*; [12]) or disruption of the information-carrying amplitude fluctuations in the target speech by the amplitude fluctuations in the masker (so-called *modulation masking*; [13]). Both energetic and modulation masking require spectro-temporal overlap between simultaneous sound sources. *Informational masking*, in contrast, refers to all the other ways background sounds interfere with speech, and it is typically associated with more central cognitive processes and higher cognitive load. The iconic informational masker is others´ speech. When listening to a specific talker amongst a background of other talkers, listeners need to engage selective attention to listen to the target talker while simultaneously ignoring all the other non-target talkers in the background. In this case, interference caused by the background speech is strongly dependent, among other factors, on the similarity of the target and masking talkers (e.g., whether of the same or different sex) and the number of talkers in the masker with greatest declines in performance observed with 1-4 talkers ([12, 14]).

Energetic/modulation masking and informational masking have different perceptual and cognitive demands, so it is not surprising that they have been shown to have a differential impact on children and older adults relative to younger adults. For example, in speech intelligibility tasks, when the interfering sound is (non-speech) noise, children achieve adult-like performance by the age of 9-10 years. However, when the interfering sound is speech, there is a prolonged developmental trajectory, with adult-like performance not reached until adolescence ([15]). Although older adults generally have greater difficulty understanding masked speech than younger adults, competing talkers seem to be particularly detrimental for older listeners as well ([16, 17, 18]). In summary, from the perspective of listening effort, background interference is more disruptive for children and older adults, especially when the interference is speech.

In terms of vocal effort, however, the age-effects associated with energetic/modulation vs. informational masking are less clear. Overall, there are well-documented changes in the acoustic-phonetic characteristics of normal everyday speech elicited in quiet listening conditions associated with age-related physiological changes. For example, children and older adults speak at a slower rate than young adults, and there are also age-related changes in fundamental frequency (f0) and the distribution of spectral energy (especially in the 1-3 kHz range) in the speech of children and older adults ([19, 20, 21, 22, 23]). In speech produced in the presence of background noise, however, the level and energetic/modulation masking potential of the noise mainly determines the amount by which vocal effort is modified and this does not differ as a function of age in adulthood ([24, 25]). Also, it has been shown that when speaking in background energetic/modulation masking noise, children and older adults make similar speech modifications to younger adults ([1,2] Smiljanic & Gilbert, 2017). Interestingly, however, a recent study by van Mersbergen et al. [26] reported that increasing the cognitive demands of a writing task led to increase in self-reported vocal effort during a reading and speaking task in adult participants. This coupling of cognitive load and vocal effort, therefore, would suggest that speech and non-speech maskers may have a differential impact on vocal effort, at least in some age groups such as children and older adults. To our knowledge, no other study has investigated the effects of energetic/modulation masking vs. informational masking on vocal effort across the lifespan.

In summary, communicative difficulty and the resulting *communication effort* exerted by a talker in background noise are likely to be modulated both by the noise type and the age of the talker via the interactive production-perception loop. Therefore, in this study we investigated, across the lifespan, the characteristics of a talker's speech when they were doing an interactive task in quiet listening conditions and when they were conversing in different types of background noise presented at levels that individuals might encounter in real-life. The following questions were addressed: i) Do children and older adults increase vocal effort to a greater extent than younger adults in background noise? ii) Does the higher cognitive load in informational relative to energetic/modulation masking lead to greater speech modifications in children and older adults relative to younger adults? iii) Does the increase in vocal effort ensure communicative success (here reflected in the completion of the interactive task)? Because of increased cognitive load associated with informational masking, we predicted that background speech would be more interfering and result in increased vocal effort as compared to non-speech backgrounds. With regards to talker age, we predict that younger groups of children and older adults would experience more interference relative to younger adults and, therefore, they need to exert more vocal effort to maintain communicative success, especially when the background noise is speech.

2. Method

2.1 Participants

A total of 114 monolingual native speakers of Standard Southern British English participated in the study, divided into six age bands: 8-12 years (M=10.3), 13-17 years (M=15.9), 18-34 years (M=21.8), 35-49 years (M=43.0), 50-64 years (M=59.3) and 65-80 years (M=71.2). Each of the age bands included 20 participants (10 female) apart from the 13-17 band due to recruitment difficulties (N=14, with only 4 males).

Participants were tested in sex- and age-group-matched pairs and were unknown to one another. They were required have a better-ear pure-tone hearing threshold average of <20 dB HL for the octave frequencies from 0.25-4 kHz. Participants reported no history of speech and language impairments or neurological trauma. All participants aged over 65 passed the Montreal Cognitive Assessment screening test (MoCA; [27]).

Ethical approval was obtained from the UCL Research Ethics Committee, and informed written consent was obtained from each participant and a parent/guardian of each child.

2.2 Procedure

The UK version of the *diapix* picture description task was used to elicit interactive semi-spontaneous speech ([28]; for a full description, see [29]). In short, *diapix* is a ´spot-the-difference´ problem-solving task where pairs of talkers work together to find differences between their pictures by conversing with each other without seeing their partner's picture ([30]). They are given 10 minutes to find the 12 differences embedded in these pictures. Each participant is assigned a role of either "lead talker" or "conversational partner". The lead talker is instructed to do most of the talking, whereas the conversational partner is mainly there to ask questions and make suggestions.

In this study, each participant carried out the *diapix* tasks as a lead talker and as a conversational partner within the same talker pair (a total of 8 different *diapix* picture pairs), and the acoustic analyses were only conducted on the speech elicited as the lead talker. The pairs took part in two test sessions, either carried out on different days or on the same day separated by a lengthy break. During each session, they also completed a series of background sensory and cognitive tasks individually (not reported in this paper). Each participant pair began with a short training to familiarize themselves with the roles of lead talker and conversational partner and with a secondary distractor task (not reported further here) in which they heard two auditory cues (a "car horn" or "dog bark") and were asked to press a bell for one and ignore the other. This secondary task aimed at increasing cognitive load and difficulty of the *diapix* task by modelling natural communicative situations that often involve conversational partners being involved in additional tasks (e.g. driving, reading, attending to media sources). Data collection took approximately 2.5 hours. Participants were paid for their participation.

Diapix was carried out in four listening conditions of which three are reported here[1]: i) quiet background ii) background noise without informational content (energetic/modulation masking: speech-shaped noise), and iii) background speech that was semantically related to the picture pair (informational masking: three voices describing the same picture). Each session started with the quiet background condition and the order of the subsequent conditions was randomised across pairs of talkers.

The informational masker consisted of 3 talkers (1 male, 1 female and a child) who were reading a pre-scripted description of each of the picture pairs (based on our previous *diapix* corpora [23]). Hesitations, false starts and speech errors were removed and the order of the description was altered for each of the masker talkers so that they were all describing a different section of the picture at any given time. The energetic/modulation masker (speech-

[1] Our initial manipulation involved two types of informational masking, one that was related to the picture-task (voices talking about the same picture) and one that was unrelated (voices talking about a different picture). We hypothesised that the semantic content of the speech masker would impact vocal effort. However, the two speech maskers yielded the same results and we therefore report results from the related-speech condition only (henceforth informational masking).

shaped noise) was created from the long-term spectrum of the recording from each voice included in the speech maskers (audio examples and experimental setup details available: https://github.com/outepi/Diapix-virtual-room).

While doing *diapix*, pairs of talkers were seated in separate acoustically-shielded rooms and they communicated via headsets fitted with a cardioid microphone (Beyerdynamic DT297). To mimic more naturalistic listening environments, we used the Spatial Audio Simulation System software for virtual rooms that simulates real room acoustics via headphones combined with head-related transfer functions in real-time (Audio 3D; available at https://www.phon.ucl.ac.uk/resource/audio3d/). The distances between the individual maskers and the "live" talkers were set to be appropriate to the virtual room dimensions (4 x 3 x 2.5 meters): the three voices (for informational masking), the equivalent three speech-shaped noise signals (for energetic/modulation masking) and the voice of the conversational partner were spatially separated 50 cm from each other and by 1 meter from the "live" talker. The intensity of the three maskers was normalised to 72 dB SPL each, and the intensity level of the "live" talkers was set to approximate a signal-to-noise ratio of 0 dB when speaking normally. Each participant was recorded on a separate channel at 44.1 kHz (16 bits) using a Fireface audio interface and Audacity audio software.

2.3 Data processing

To measure task difficulty (as a proxy for "communication success"; [10]), we calculated the average time it took for each talker pair to identify one difference (transaction time; Time-1); that is, the total time spent on the task divided by the number of differences found. For acoustic analyses, all recordings were semi-automatically transcribed using a cloud-based transcription system (https://www.speechmatics.com/) and manually corrected in Praat ([31]). Three acoustic-phonetic measures were extracted from the lead talker´s audio recordings: articulation rate, median f0 and mean energy in the mid-frequency range (1-3 kHz) of the long-term average spectrum. These three measures have been previously shown to reflect age-related changes in acoustic characteristics of speech ([19,20, 22]), and they have been associated with different intelligibility-enhancing speaking styles such as clear speech, Lombard speech and shouted speech ([2, 3]). In this study, the definition and measure of vocal effort is based on our previous work where we report that, when communication is made difficult, increases in mid-frequency energy is coupled with increases in median f0 ([22,23]). Therefore, vocal effort is operationalised here as a positive correlation between the median f0 and mid-frequency energy measures. As talkers vary considerably in their acoustic profiles in casual everyday speech elicited in quiet, values are reported as difference scores (NOISE − QUIET) in order to get a measure of speech modifications in challenging conditions.

For further details regarding acoustic measures, see [23]. In short, articulation rate was calculated as the number of syllables divided by the total duration in seconds of the speech regions, thus excluding pauses, hesitations and fillers. For median f0, all speech was first concatenated and f0 calculations were then done in Praat ([31]) using the “pitch” function with a time step of 100 pitch values per second and converted to semitones. Because estimates of f0 are prone to error, we first set limits on the range of frequencies considered valid for each talker based on the 25th and 75th percentiles before calculating medians ([32]). The mid-frequency spectral energy was measured in Praat by first calculating the intensity of labelled speech segments and excluding values above 88 dB as likely instances of shouting. The remaining segments were concatenated and scaled to the arbitrary Praat-referenced level of 75 dB. The signal was then band-pass filtered between 1 and 3 kHz and the mean intensity of the resulting waveform calculated to give a measure of the amount of energy in the 1–3 kHz frequency range relative to the total energy in the spectrum. This particular mid-frequency band was chosen because the 1-3 kHz region is rich in acoustic cues to phonetic identity and an increase in the relative energy in this range reflects a reduction of spectral tilt which has been documented in speech produced with greater vocal effort (e.g., [33]).

3. Data analysis and results

First, the quiet background condition was analysed using a between-subjects anova to investigate the effects of talker age (8-12 years, 13-17 years, 18-34 years, 35-49 years, 50-64 years, 65-80 years) and sex (F,M) on the acoustic characteristics of spontaneous speech (articulation rate, median f0, mid-frequency energy) produced in good listening conditions. Because of well-known acoustic differences between male and female speech, sex was also used a predictor in the analyses.

Second, speech adaptations in energetic/modulation and informational masking (relative to the quiet background) were analysed with a series of mixed-effects models using the lme4 package of R ([34]). Models were constructed separately for each outcome measure (articulation rate, median f0, mid-frequency energy, transaction time) but were otherwise set-up identically. Participant was set as a random intercept, with fixed effects of talker age, sex and listening condition (energetic/modulation masking, informational masking), and all their interactions. Predictors were then removed iteratively if their removal did not significantly worsen the fit of the model.

Third, vocal effort was quantified as the Pearson correlation between the median f0 and mid-frequency energy (both relative to the quiet background) separately for each age group.

Pseudo R^2 (Nagelkerke) was used as an index of goodness-of-fit. All follow-up tests were conducted with the estimated marginal means package in R (emmeans), Tukey´s test and t-tests. Data and analysis scripts for R are available at https://github.com/outepi/Speech-modifications-data.

3.1 Speech produced in a quiet background

As expected based on several earlier studies, for articulation rate, there were significant main effects of age group [F(5, 102)=8.55; p<.001] and sex [F(1, 102)=4.25; p=.042] and no significant interactions. Post-hoc analyses (Tukey´s test) showed that the youngest (8-12 years: M=3.79 s/s) and the two oldest groups (50-64 years: M=4.14 s/s; 65-80 years: M=3.92 s/s) had a slower articulation rate than 13-49 year olds (13-17 years: 4.31 s/s; 18-34 years: 4.46 s/s and 35-49 years: 4.16 s/s; all comparisons p≤.009). Females (M=4.10 s/s) spoke slower than males (M=4.25 s/s).

As regards fundamental frequency, there were, as expected, main effects of age group [F(5,102)=68.22; p<.001], sex [F(1,102)=196.24; p<.001] and a sex by age group interaction [F(5,102)=6.97; p<.001] (see Fig 1). Median f0 for the 8-12 year olds was higher than all other age groups (all comparisons, p<.001), and it was also higher for the 13-17 year olds than 35-49 year and 50-64 year olds (p=.001 and p=.033 respectively). The interaction between age group and sex was due to the lack of a sex-related difference in the 8-12 year olds only (t=1.16.; p=.261; see Fig. 1).

For mid-frequency energy, there were again effects of age group [F(5, 102)=6.14; p<.001) and sex [F(1, 102)=29.05; p<.001) and no significant interactions. The 8-12 year olds had higher mid-frequency energy (M=67.1 dB; all comparisons, p≤.007) than the other five age groups that did not differ (Means: 63.5, 63.6, 64.3, 63.4, 64.1 dB, respectively; p≥.859). Mid-frequency energy was higher in female (M=65.5 dB) than in male speech (M=63.1 dB).

Overall, our results for speech produced in quiet background replicate previously reported age- and sex-related changes: adult-like speaking rate emerges at around 13 years of age and slows down again after 50 years of age, and median f0 and both decrease at around 13 years of age with notable sex-related differences in f0 in this timeframe.

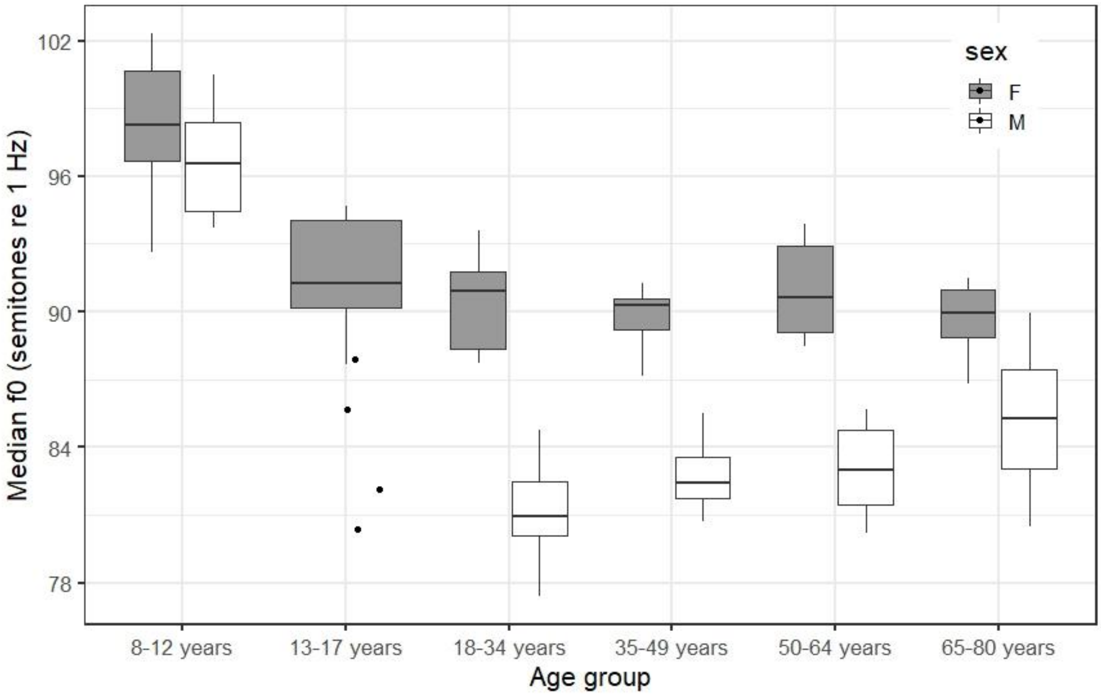

Fig 1. Median f0 (in semitones relative to 1 Hz) for the six age groups and for female and male talkers separately in the quiet background condition. The 13-17 year male group has N=4 and individual data points are shown (M=84.1). Error bars indicate 95% confidence intervals.

3.2 Speech adaptations in challenging listening conditions

We then investigated whether the age groups differed in the adaptations made as reflected in the acoustic measures when communicating in challenging conditions (informational masking and energetic/modulation masking relative to the quiet background). Reducing articulation rate was a strategy used for both noise conditions and there was no significant effect of condition or interactions with it (see Fig. 2). The best fitting model included significant main effects of sex [$\chi 2$ (1)=6.35, p=.012, R^2_m=.047] and age group [$\chi 2$ (5)=11.29, p=.046, R^2_m=.121] and their two-way interaction [$\chi 2$ (5)=24.38, p<.001, R^2_m=.255]. Post hoc comparisons revealed that there were sex-related differences only in the 8-12 year olds (t=-3.52, p=.001), the 18-34 year olds (t=2.11, p=.037) and the 50-64 year olds (t=-3.78, p<.001) showing that, even though women were slower speakers in the quiet background condition, they reduced their articulation rate to a greater extent than men, apart from the 18-34 year olds where the pattern was reversed.

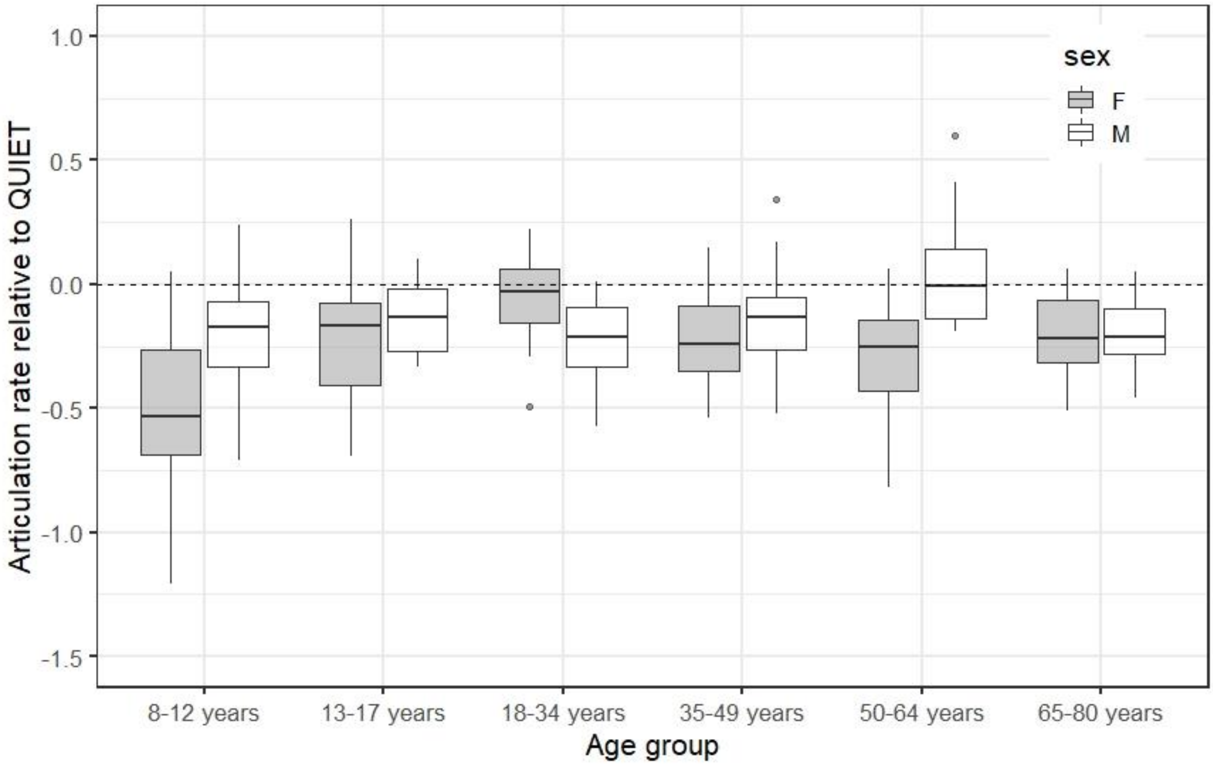

Fig 2. Articulation rate (aggregated over noise type) relative to the quiet condition for the six age groups and the female and male talkers. Negative values indicate slower speaking rate for noise than for quiet background and dashed line at 0 indicates no difference between the noise and quiet conditions. Error bars indicate 95% confidence intervals.

For fundamental frequency, the best fitting model included significant main effects of sex [χ2 (1)=19.75, p<.001, R^2_m=.135], condition [χ2 (1)=56.45, p<.001, R^2_m=.191], age group [χ2 (5)=16.42, p=.006, R^2_m=.278] and an interaction between condition and age group [χ2 (5)=18.28, p<.001, R^2_m=.287]. Overall, there was a greater increase in median f0 for males (3.9 st) than females (2.8 st) and for energetic/modulation masking (3.6 st) than for informational masking (3.0 st) (see Fig. 3). Post hoc comparisons for the condition and age group interaction revealed that all groups, apart from 13-17 year olds (p=.121), increased their f0 more for the energetic/modulation masking condition than for the informational masking condition (all comparisons, p≤.025). Furthermore, within the energetic/modulation masking condition, the 8-12 year (p<.001), 35-49 year (p=.028) and 50-54 year olds (p=.017) groups increased their median f0 more than the 13-17 year olds. The age groups did not differ in the informational masking condition (p≥.081).

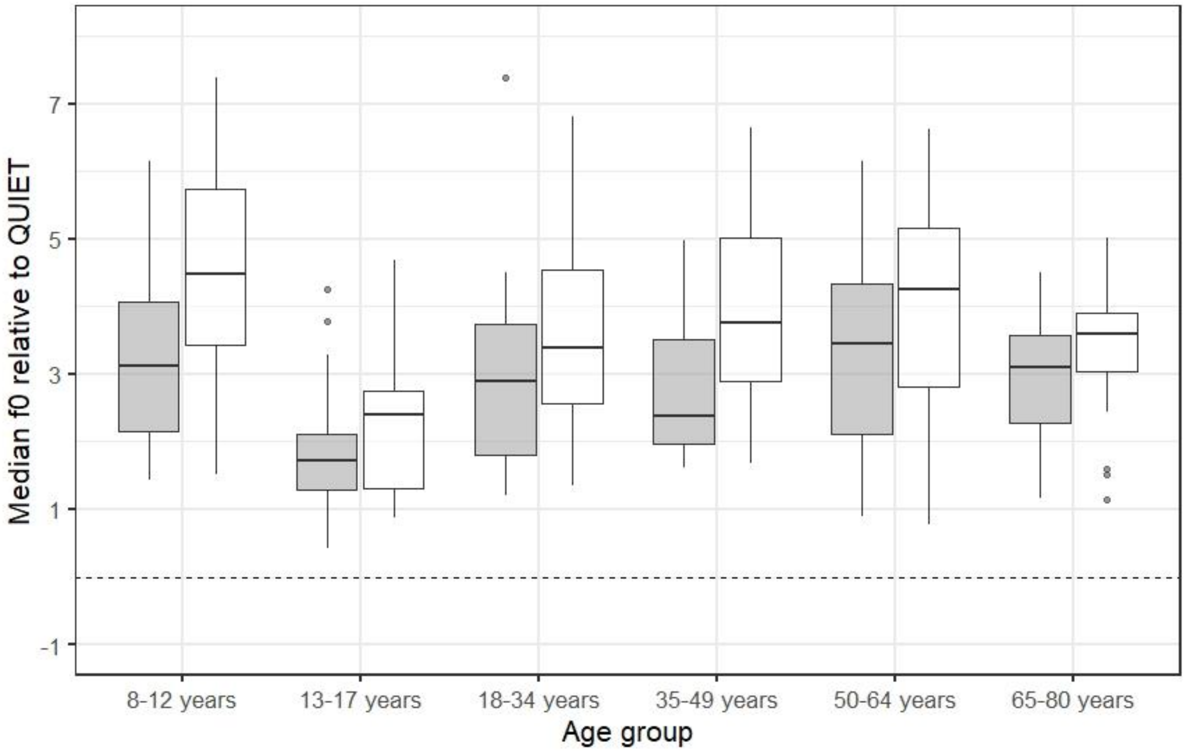

Fig 3. Median f0 change (in semitones) for the six age groups (aggregated over sex) and for the noise relative to quiet condition (energetic/modulation masking condition in white and informational masking condition in grey). Dashed line at 0 indicates no difference between the noise and quiet conditions. Error bars indicate 95% confidence intervals.

For mid-frequency energy, the only significant effect was that of condition [χ2 (1)=48.52, p<.001, R^2_m=.021] with a greater increase in energetic/modulation masking (M=3.4 dB) than informational masking condition (M=2.9 dB).

3.3 Vocal effort in challenging listening conditions

The next analysis investigated whether there was evidence of vocal effort, shown by a correlation between the f0 and mid-frequency energy, as found in previous studies for groups likely to be experiencing greater interference in background noise, such as children and older adults (Hazan et al., 2016; Hazan et al., 2018). Here, for energetic/modulation masking, correlations were only significant for the 50-64 year and 65-80 year olds, explaining 26% and 34% of the variance. For the informational masking condition, typically associated with greater cognitive load, correlations were again significant for those two older groups (50-64 year and 65-80 year olds) and also for the youngest group of children (8-12 years) explaining 23-41% of the variance (see Table 1).

Table 1: Correlations between the median f0 (semitones) and mid-frequency energy (in dB) (relative to the quiet background condition) for the energetic/modulation masking and informational masking conditions for the six age groups. Significant effects highlighted in bold font.

	Energetic/modulation masking		**Informational masking**	
	r	**p**	**r**	**p**
8-12 years, *N=20*	0.01	.966	**0.54**	**.013**
13-17 years, *N=14*	-0.05	.862	-0.31	.284
18-34 years, *N=20*	-0.02	.933	-0.02	.924
35-49 years, *N=20*	0.16	.506	-0.09	.717
50-64 years, *N=20*	**0.58**	**.008**	**0.64**	**.002**
65-80 years, *N=20*	**0.51**	**.021**	**0.48**	**.033**

3.4 Task difficulty

In order to assess communicative success, the Time-1 communicative efficiency measure was analysed for each participant pair in each condition (quiet, energetic/modulation masking, informational masking).

The best fitting model included significant main effects of age group [χ2(5)=25.53, p<.001, R^2_m=.169] and condition [χ2(2)=2.88, p=.237, NS] and the interaction between age group and condition [χ2(10)=20.90, p=.022, R^2_m=.171]. The youngest children (8-12 years: M=63.6 sec) took longer to find differences overall but the other groups did not differ in their task transaction time [13-17 years (M=43.7 sec), 18-34 years (M=40.7 sec), 35-49 years (M=46.5 sec), 50-64 years (M=48.0 sec) and 64-80 years (M=49.4 sec)]. Post-hoc tests carried out per age group showed that there was no effect of condition for any of the groups.

In summary, all six groups of participants managed to find differences as quickly in the masking conditions as when carrying out the task in quiet. This lack of listening condition difference in the Time-1 measure that suggests the acoustic adaptations in background noise were successful in maintaining efficient communication.

4. Discussion

For speech communication in good listening conditions, we found age- and sex-related patterns that can be linked with physiological changes, as reported in several previous studies ([19, 20, 21]). When communication was challenging we expected to see differences related to the background noise type (energetic/modulation masking vs. informational masking) and talker age. Our results showed that speaking slower was a strategy applied by all groups, regardless of the background noise type. Notably, however, the amount of reduction in speaking rate in background noise was small. The just noticeable difference for perceptual changes in speech tempo in read speech is approximately 5% ([35]), and only the youngest group of children (8-12 years) and oldest group of adults (65-80 years) decreased their articulation rate beyond the threshold where the change in articulation rate is perceptually salient. Another temporal strategy that the talkers might have used is to increase pausing but this was not measured in the current study.

Furthermore, all groups increased the median f0 (apart from the 13-17 year olds) and mid-frequency energy in their voice more for energetic/modulation masking than for informational masking. Although, as with articulation rate, the increase in mid-frequency energy is small and unlikely to be audible to the listener. These results are partially in-line with previous literature on Lombard speech that have shown that the energetic/modulation masking potential of the noise mainly determines the amount by which vocal adaptations are applied, and that this does not change as a function of age ([24,25]).

However, the most interesting findings were related to vocal effort (as marked by correlated changes in f0 and mid-frequency energy), which could be seen as an indication of higher perceptual interference (energetic/modulation masking) and greater cognitive load (informational masking) imposed by the background noise type. We expected that the presence of (meaningful) background competing speech creates the need to allocate extra attentional resources to focus on the target speaker and ignore the background distractor voices, and the increased cognitive load would in turn lead to increases in vocal effort ([26]). Our results showed that the two older groups (50-80 year olds) increased vocal effort in the presence of both energetic/modulation and informational masking, and the youngest group of children (8-12 year olds) showed this pattern in the presence of informational masking only. Other age groups did not show increased vocal effort, at least as has been defined here.

We have shown similar age-related effects, using the same measure, in our previous work on clear speech strategies, in that children (9-14 years) and older adults with age-related hearing loss (65-85 years) tend to primarily increase vocal effort while younger adults vary in the strategies they adopt to increase the intelligibility of their speech ([22,23]). The fact that, here, we see increased vocal effort already in the 50+ age range and in individuals with normal hearing is likely to do with the differences in experimental set-ups. In our previous studies, communication between the dyads was made maximally difficult by simulating the effects of severe-to-profound hearing loss or a cochlear implant in the conversational partner while the lead talker was hearing their partner normally. In the current study we placed both members of the pair in background noise, and although the overall level of the noise is relatively moderate, it can impose different perceptual and cognitive demands for younger children and older adults regardless of hearing acuity. For example, it has been shown that both children and older adults have greater difficulty understanding masked speech than younger adults, with age-related differences in the adult lifespan emerging already in midlife (see [36] for a review).

With regards to communicative success, we found that there were no differences in how quickly the dyads found the differences among any of the communicative conditions~~.~~ Overall, these findings suggest that whether background noise is disrupting more peripheral auditory function (as in the case of energetic/modulation maskers) or whether it has a greater cognitive load (as in the case of informational maskers), age groups aged 50 years and above need to exert more vocal effort in order to ensure successful communication. However, it is only in the case where the background noise has a higher cognitive load that children aged 8-12 years needed to do so to successfully complete the task.

In this study, we focused on vocal effort and investigated acoustic-phonetic modifications made by the “lead talker” when communication was made challenging. As such, our results might not be fully reflective of real-life communicative situations, as exemplified by the *communication effort* framework ([10]). For example, in communicative settings, the roles of a “talker” and a “listener” are often fluid. When interacting with another person, especially in challenging conditions, talkers are constantly monitoring (i.e., listening to) their own speech as well as feedback from their conversational partner (the listener) in order to optimise communicative success. This enables them to quickly react to changes in the external and internal perceptual demands and to increase/decrease investment of vocal and listening effort according to these demands. At the same time, the listener is constantly monitoring the incoming speech while planning and executing a verbal response. Therefore, both the talker and the listener are simultaneously engaging resources for vocal effort and listening effort, which are also prone to fluctuate as a function of "communicative success" during the course of a conversation. For example, in the current study, the lack of evidence for vocal effort in some of the groups does not necessarily mean effortless communication, as the other groups may have invested in increasing listening effort instead. Partial evidence for this comes from the ratings obtained from adults over the lifespan (18-80 years) that showed that, generally, the presence of energetic/modulation maskers was perceived as less effortful, requiring less concentration and easier to ignore than the informational masking condition by all groups of adult participants ([37]).

Furthermore, the communicative tasks and measures of communicative success used in these laboratory-based experiments do not necessarily correspond to demands imposed by everyday conversations. For example, our results showed increased vocal effort in some groups of participants when engaged in the problem-solving *diapix* task while also attending a secondary task. Increasing vocal effort in order to finish the *diapix* task faster is a strategy that works well when the goal and success of the task is well defined ("find 12 differences") and the task needs to be completed within a particular timeframe (here, 10 minutes). Other similar studies have used tangram and Sudoku puzzles that all have clear goals and outcome measures of communicative success ([10,24]). Most of our everyday interactions are rarely like this: we usually do not have an immediate goal that can be measured in terms of success and the duration of our conversations are often unpredictable, or at least are not pre-set. Increasing vocal effort to improve intelligibility for longer or unknown periods of time is a strategy that can lead to both vocal and psychological fatigue which, again, can be detrimental to communicative success. Therefore, it is feasible that in some more realistic situations, talkers might choose to sacrifice the efficiency of message transmission to prevent fatigue. In that case we would expect intelligibility levels of the speech to fluctuate and vocal effort to remain constant. Furthermore, communicative success in real life is not necessarily something we can measure at the end of a conversation. Rather, it is an ongoing fluctuation between utterances understood and misunderstood.

Finally, defining and quantifying “vocal effort” has proven to be as elusive as quantifying listening effort. The term “vocal effort” is often used interchangeably with the terms “speaking effort”, “vocal fatigue” and “vocal load” with inconsistencies between studies over how these are measured. In an attempt to develop a consensus over terminology and definitions, Hunter et al., [5] proposed a distinction between perceptual and physiological manifestations of effortful speaking that resembles the framework proposed for effortful listening (FUEL). In their framework, “vocal effort” refers to the increased contextual demands experienced by the talker (e.g., background noise) which is often followed by the “vocal demand response” (e.g., clear speech and Lombard speech adaptations). According to Hunter et al. ([5]), vocal effort as a more perceptual phenomenon is best assessed by talker self-reports whereas the vocal demand response can be directly measured from the talker´s voice characteristics or the acoustic-phonetic characteristics of their speech.

It is noteworthy, however, that vocal effort is multidimensional and context-dependent, and it can involve a continuum of speaking styles with varying degree of voluntary control (e.g., clear speech, Lombard speech, shouted speech). Here, for example, the dyads were matched on age and sex in order to elicit speech that is close to casual everyday speech and to minimise effects of phonetic convergence in measures where we have clear sex-related effects. Pairing individuals from different age bands (e.g., a younger adult with a child or with an older adult) might elicit additional speech styles (to Lombard speech) such as child-directed speech or `elderspeak´ as in the case of younger adult –older adult couplings, that would confound the effects of background noise type and talker age on vocal effort. Therefore, it is unlikely that a change in single acoustic dimension or in a combination of dimensions can fully capture vocal effort across these different conversational contexts.

In the current study, in line with previous work in interactive speech and speech produced in challenging conditions, we operationalised vocal effort as concomitant changes in mid-frequency energy and f0 ([3, 22, 23]). Other measures to capture vocal effort in different contexts might include acoustic perturbation measures associated, for example, with prolonged voice use (e.g., jitter, noise-to-harmonic ratio), f0 range and variability, segmental hyperarticulations (vowels, consonants) and perceptual ratings and evaluations of voice.

Moreover, operationalising effortful speaking and effortful listening according to the dichotomy of perceived and expressed effort is not straightforward. While subjective effort ratings have been successfully used in speech-in-noise perception experiments ([38]) and in spoken conversations between conversational dyads ([10, 37]), individuals can differ greatly how they conceptualise "effort", especially when comparing children and adults. A more promising avenue for quantifying the cognitive demands associated with listening and speaking effort in laboratory settings might arise from dual-tasking paradigms where both the cognitive load of the secondary task and the consequent physiological manifestations of speaking and listening effort and its consequences for task success can be better operationalised. In summary, in order to quantify effort experienced by two or more individuals engaged in a conversation, both cognitive effort (e.g., self-reports and dual-task performance) and physiological effort associated with listening and speaking (e.g., pupillometry, acoustic-phonetic characteristics of speech) should be measured as a function of an instantaneous measure of communicative success (e.g., after a misunderstanding/feedback signalling understanding).

To conclude, the results from the current study demonstrate that the both talker age and the type of background noise impacts the vocal modifications talkers make in order to increase the intelligibility of their speech. As suggested by Beechey et al. ([10]) and Hunter et al. ([5]), our study also highlights the need for a unified framework for vocal and listening effort in order to account for the communicative difficulties experienced in everyday conversational settings.

Acknowledgements

This work was supported by the United Kingdom Economic and Social Research Council (Grant No. ES/P002803/1).

References

[1] Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

[2] Smiljanić, R., & Bradlow, A. R. (2008). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Language and Linguistics Compass,* 3(1), 236–264. DOI: https://doi.org/10.1111/j.1749-818x.2008.00112.x

[3] Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language, and Hearing Research*, 53(3), 588–608. DOI: https://doi.org/10.1044/1092-4388(2009/08-0138)

[4] Junqua, J. C., Finckle, S., & Field, K.(1999). The Lombard effect: A reflex to better communicate with others in noise. In Proceedings of ICASSP'99, the International Conference on Acoustics, Speech and Signal Processing (pp. 2083–2086). DOI: 10.1109/ICASSP.1999.758343

[5] Hunter, E. J., Cantor-Cutiva, L. C., van Leer, E., van Mersbergen, M., Nanjundeswaran, C. D., Bottalico, P., Sandage, M. J., & Whitling, S. (2020). Toward a Consensus Description of Vocal Effort, Vocal Load, Vocal Loading, and Vocal Fatigue. *Journal of Speech, Language, and Hearing Research*, 63(2), 509–532. DOI: https://doi.org/10.1044/2019_jslhr-19-00057

[6] Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear & Hearing*, 37(1), 5S-27S. DOI: https://doi.org/10.1097/aud.0000000000000312

[7] McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper.' *International Journal of Audiology*, 53(7), 433–445. DOI: https://doi.org/10.3109/14992027.2014.890296

[8] Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. The Journal of the Acoustical Society of America, 141(6), 4680–4693. https://doi.org/10.1121/1.4986938

[9] Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear & Hearing*, 39(2), 204–214. DOI: https://doi.org/10.1097/aud.0000000000000494

[10] Beechey, T., Buchholz, J. M., & Keidser, G. (2020). Hearing Impairment Increases Communication Effort During Conversations in Noise. *Journal of Speech, Language, and Hearing Research*, 63(1), 305–320. DOI: https://doi.org/10.1044/2019_jslhr-19-00201

[11] Lindblom B. (1990). “Explaining phonetic variation: a sketch of the H&H theory,” in Speech Production and Speech Modelling eds. Hardcastle W. J., Marchal A. (New York, NY: Springer;) 403–439. DOI: 10.1007/978-94-009-2037-8-16

[12] Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5), 182–186. DOI: https://doi.org/10.1016/j.tics.2008.02.003

[13] Stone, M. A., Füllgrabe, C., & Moore, B. C. J. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, 132(1), 317–326. DOI: https://doi.org/10.1121/1.4725766

[14] Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, 133(4), 2431–2443. DOI: https://doi.org/10.1121/1.4794379

[15] Leibold, L. J., & Buss, E. (2019). Masked Speech Recognition in School-Age Children. *Frontiers in Psychology*, 10. DOI: https://doi.org/10.3389/fpsyg.2019.01981

[16] Goossens, T., Vercammen, C., Wouters, J., & van Wieringen, A. (2017). Masked speech perception across the adult lifespan: Impact of age and hearing impairment. *Hearing Research*, 344, 109–124. DOI: https://doi.org/10.1016/j.heares.2016.11.004

[17] Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear and hearing*, 29(1), 87–98. DOI: https://doi.org/10.1097/AUD.0b013e31815d638b

[18] Schoof, T., & Rosen, S. (2014). The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners. *Frontiers in Aging Neuroscience*, 6. DOI: https://doi.org/10.3389/fnagi.2014.00307

[19] Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2), 839–850. DOI: https://doi.org/10.1121/1.3459842

[20] Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., & van Lieshout, P. (2013). Normative Voice Data for Younger and Older Adults. *Journal of Voice*, 27(5), 545–555. DOI: https://doi.org/10.1016/j.jvoice.2013.03.002

[21] Schaeffer, N., Knudsen, M., & Small, A. (2015). Multidimensional Voice Data on Participants With Perceptually Normal Voices From Ages 60 to 80: A Preliminary Acoustic Reference for the Elderly Population. *Journal of Voice*, 29(5), 631–637. DOI: https://doi.org/10.1016/j.jvoice.2014.10.003

[22] Hazan, V., Tuomainen, O., & Pettinato, M. (2016). Suprasegmental Characteristics of Spontaneous Speech Produced in Good and Challenging Communicative Conditions by Talkers Aged 9–14 Years. *Journal of Speech, Language, and Hearing Research*, 59(6). DOI: https://doi.org/10.1044/2016_jslhr-s-15-0046

[23] Hazan, V., Tuomainen, O., Tu, L., Kim, J., Davis, C., Brungart, D., & Sheffield, B. (2018). How do aging and age-related hearing loss affect the ability to communicate effectively in challenging communicative conditions? *Hearing Research*, 369, 33–41. DOI: https://doi.org/10.1016/j.heares.2018.06.009

[24] Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskersa). *The Journal of the Acoustical Society of America*, 128(4), 2059–2069. DOI: https://doi.org/10.1121/1.3478775

[25] Smiljanić, R., & Gilbert, R. C. (2017). Acoustics of Clear and Noise-Adapted Speech in Children, Young, and Older Adults. *Journal of Speech, Language, and Hearing Research*, 60(11), 3081–3096. DOI: https://doi.org/10.1044/2017_jslhr-s-16-0130

[26] van Mersbergen, M., Vinney, L., & Payne, A. (2020). Cognitive influences on perceived phonatory exertion using the Borg CR10. *Logopedics PhoniatricsVocology*, 45:3, 123-133. DOI: 10.1080/14015439.2019.1617895

[27] Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 53, 695–699.

[28] Baker, R., & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43, 761–770. DOI: 10.3758/s13428-011-0075-y.

[29] Tuomainen, O., & Hazan, V. (2018). Investigating Clear Speech Adaptations in Spontaneous Speech Produced in Communicative Settings. In In: Challenges in Analysis and Processing of Spontaneous Speech. MTA Nyelvtudományi Intézet. DOI: https://doi.org/10.18135/CAPSS.9

[30] Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Lang. Speec,* 53, 510–540.

[31] Boersma, P. & Weenink, D. 2018. Praat: doing phonetics by computer. http://www.praat.org/

[32] De Looze, C. & Hirst, D. J. 2008. Detecting key and range for the automatic modelling and coding of intonation, *Actes de Speech Prosody*, Conference, Campinas, 135-138.

[33] Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America,* 100, 2471– 2485.

[34] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. DOI: 10.18637/jss.v067.i01.

[35] Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3), 353–362. https://doi.org/10.1016/j.wocn.2006.09.001

[36] Helfer, K. S., & Jesse, A. (2021). Hearing and speech processing in midlife. Hearing Research, 402, 108097. https://doi.org/10.1016/j.heares.2020.108097

[37] Hazan, V., Tuomainen, O. & Taschenberger, L. (2019) Subjective Evaluation of Communicative Effort for Younger and Older Adults in Interactive Tasks with Energetic and Informational Masking. Proc. Interspeech 2019, 3098-3102, DOI: 10.21437/Interspeech.2019-2215

[38] Zekveld, A., Rudner, M., Kramer, S. E., Lyzenga, J., & Ronnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masked speech. *Frontiers in Neuroscience,* 8, 88. doi 10.3389/fnins.2014.00088.