ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 98 (2012) 317 – 327

DOI 10.3813/AAA.918515

The Relationship Between Concurrent Speech Segregation, Pitch-Based Streaming of Vowel Sequences, and Frequency Selectivity

Etienne Gaudrain[1)]*, Nicolas Grimault[1)], Eric W. Healy[2)], Jean-Christophe Béra[3)]

[1)] Cognition Auditive et Psychoacoustique, Lyon Neuroscience Research Center, CNRS 5292, Inserm 1028, Université de Lyon, Lyon, France. ngrimault@olfac.univ-lyon1.fr

[2)] Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio, USA

[3)] Laboratory of Therapeutic Applications of Ultrasound, Inserm 1032, Lyon, France

Summary

Simultaneous and sequential segregation form the basis of auditory scene analysis and are likely involved in concurrent speech segregation. However, previous work showed that speech-in-noise perception was uncorrelated with simultaneous segregation, whereas it appeared to be related to the pure-tone fusion threshold of sequential streaming. The current study aimed to clarify the relationships between pitch-based speech-in-speech segregation, pitch-based streaming, and frequency selectivity. Twenty-three listeners with close to normal hearing were involved. Speech-in-speech perception was measured using words presented in a time-reversed single talker background, with various pitch differences between target and masker. Streaming performance was measured using an objective order-naming task on vowel sequences. Auditory filter widths were derived using a notch-noise method. Results showed a correlation between the effect of pitch on speech-in-speech perception and the effect of pitch on streaming performance. However, frequency selectivity was found to correlate with average speech-in-speech perception but not with streaming, and only in the region of the second formant. These latter results are consistent with the hypothesis that pitch-based streaming probably relies on pitch discrimination, which is only poorly correlated to frequency selectivity. Further, these results suggest that mild impairments in frequency selectivity do not systematically impair pitch-based streaming.

PACS no. 43.66.Mk, 43.66.Sr, 43.71.Es, 43.71.Ky

1. Introduction

Simultaneous and sequential segregation are commonly considered base mechanisms of Auditory Scene Analysis (ASA) [1], and are certainly involved in the resolution of Cocktail Party situations [2] or in speech-in-speech perception tasks. This relationship has motivated the search for correlations between simultaneous or sequential segregation and speech perception performance. The perception of speech in a masker and the mechanisms of ASA seem to be subject to a number of common factors, among which pitch is probably one of the most important.

Brokx and Nooteboom [3] reported that a pitch difference between two sentences uttered by the same speaker increased intelligibility of the target sentence. A pitch difference as small as 3 semitones was enough to increase correct responses by 20%. Similarly, Summers and Leek [4] found an improvement of more than 10% in normal-hearing (NH) listeners when adding a pitch difference of 4 semitones between simultaneous synthetic sentences. In these two reports, the percentage of correct responses grew roughly linearly with the pitch difference in semitones. Bird and Darwin [5] used a speech masker that was almost entirely voiced to enhance the effect of F_0. They observed an increase in word recall of 40% between 0 and 2 semitones, and an additional increase of 20% between 2 and 8 semitones. More recently, Darwin *et al.* [6] used concurrent sentences from the Coordinate Response Measure speech corpus to observe the effect of a difference in F_0 and in vocal tract length. They observed that the reception score for the target sentence increased by 24% between 0 and 12 semitones.

In contrast, the benefit of F_0 difference (ΔF_0) for concurrent vowel identification is saturated over 2 semitones [7, 8]. This difference in the range of ΔF_0 over which concurrent sentence and vowel identification improves suggests that simultaneous segregation is not the only pitch-based segregation mechanism involved in concurrent sentence perception. Summers and Leek [4] highlighted this difference by comparing performances in concurrent-vowel and concurrent-sentence tasks in NH and hearing-

Received 15 April 2011,
accepted 2 January 2012.

* Currently at the MRC Cognition and Brain Sciences Unit, Cambridge, UK. etienne.gaudrain@mrc-cbu.cam.ac.uk

impaired (HI) listeners. These authors found that the F_0-related benefit in the concurrent-sentence task was not clearly associated with the F_0-related benefit in the concurrent-vowel task, especially in HI listeners.

Early reports involving pure/complex tones indicated that streaming can be induced over a wider range of $\Delta F/\Delta F_0$'s relative to simultaneous segregation [9]. In a more recent study, the effect of ΔF_0 on the streaming of synthesized vowel sequences was found to grow continuously from 0 to 12 semitones [10], a range close to the ΔF_0 benefits observed for concurrent speech perception. In line with this observation, a few studies investigated the potential of sequential segregation as a predictor of speech-in-speech perception. Mackersie *et al.* [11] studied the relationship between streaming and performance in a concurrent-sentence task in NH and HI listeners. Streaming was evaluated using the fission threshold for tones differing in frequency up to 6 semitones. The fission threshold is defined as the frequency difference below which a tone sequence can no longer be perceived as two streams and is instead perceived as a single stream (see [12] for details). Concurrent sentence recognition involved sentence pairs produced by one female talker (mean $F_0 = 240\,\text{Hz}$) and one male talker (mean $F_0 = 115\,\text{Hz}$). The results revealed that sequential segregation and concurrent speech perception were strongly correlated. Hong and Turner [13] also observed such a correlation in cochlear implant users between streaming of pure tones and speech perception in steady-state noise and multi-talker babble.

Pitch perception and auditory tuning is often impaired in HI listeners [14]. The effect of frequency selectivity on speech-in-noise has been clearly established when the masker is steady-state or amplitude-modulated noise [15, 16, 17, 18]. It has also been found that HI listeners benefit less from F_0 differences in concurrent-sentence tasks than do NH listeners (for $\Delta F_0 \geq 4$ semitones) [4]. However, no relationship was found between ΔF_0 benefit and frequency selectivity [11]. It is therefore possible that frequency selectivity influences speech perception in noise but does not influence speech in speech segregation based on ΔF_0.

The results are also not entirely consistent when the relation between tuning and streaming is examined. Rose and Moore [12] investigated the sequential segregation of pure tone sequences in unilaterally-impaired listeners, but found no clear relationship between auditory filter width and fission boundary. In contrast, Grimault *et al.* [19] observed that hearing impairment had an influence on F_0-based streaming when this impairment affected the resolvability of the harmonics composing the complexes. More recently, Gaudrain *et al.* [10] observed a significant deficit in streaming of vowel sequences when moderate to severe auditory filter broadening was simulated using the algorithm developed by Baer and Moore [17]. However, frequency selectivity was reduced using a single smearing value by Gaudrain *et al.*, which cannot account for the normal variations in selectivity that can affect segregation in the population.

The purpose of the current study was to clarify the relations between ΔF_0 benefit in speech-in-speech perception, F_0-based streaming with vowels, and frequency selectivity. Variation in frequency selectivity was obtained by selecting listeners having normal to slightly-impaired hearing. The use of subjects having mild impairments allowed an examination of auditory filter variation without concerns involving audibility. Each volunteer participated in three tasks: speech-in-speech reception was measured using word lists in a reversed speech background, streaming was evaluated using an objective order-naming task on vowel sequences [10], and auditory filter width was evaluated using a notched-noise method.

2. Method

2.1. Listeners

Twenty three listeners with normal to slightly impaired hearing participated. They ranged in age from 18 to 27 years, with a mean age of 21 years. Listeners were selected on the basis of their audiometric thresholds. To simplify the selection of listeners and specification of auditory tuning, they were tested in only one ear. Their audiometric thresholds [20] in the test ear are shown in Table I. Listeners were paid an hourly wage for their participation and provided informed consent.

2.2. Procedure

The order of the three tasks was randomized for each subject. Subjects completed one task prior to moving to the next, and returned to participate in the different tasks on different days. All stimuli were presented using a PC, a Digigram VxPocket 440 soundcard, a Behringer Ultragain amplifier and a Sennheiser HD250 Linear II headphone. Sound levels were calibrated in an artificial ear (Larson Davis AEC101 and 824, [21]). All the experiments took place in a sound attenuated booth. The experimental procedure was formally approved by a local ethics committee (CPP Lyon Sud).

2.3. Frequency selectivity

2.3.1. Materials

Auditory filters were derived using a symmetric notched-noise masker and sinusoidal probe tone [22, 23], with fixed probe level [24]. Auditory filter width was measured for two frequencies: 370 Hz and 1394 Hz. These correspond to the average frequencies of the first and second formant over all six vowels used in the streaming test described in the next section. The masker consisted of a white noise in which a notch was created using a 16th-order Butterworth band-stop filter. Cutoff frequencies for the notch are expressed as a proportion r of the center frequency f_c as follows: $(1-r)f_c$ and $(1+r)f_c$, thus forming a symmetric notch on a linear frequency scale. In addition, the signal was bandpass filtered between $0.2f_c$ and $1.8f_c$ with a 4th order Butterworth filter, in order to maintain the overall level relatively low while ensuring proper masking of the tone. Finally a lowpass noise was added below $0.2f_c$ (4th order Butterworth filter), at a level 20 dB below that

Table I. Audiometric thresholds in dB HL of the 23 listeners involved in the study. Test ear is indicated in the second column.

		Frequency (Hz)						
Listener	Test ear	250	500	1000	2000	4000	6000	8000
S01	R	20	20	10	0	0	15	5
S02	R	20	15	0	10	15	15	10
S03	L	20	10	5	10	30	40	5
S04	L	20	25	15	20	5	35	30
S05	L	25	25	25	5	10	10	5
S06	L	20	20	5	10	5	15	5
S07	L	15	10	10	5	10	10	10
S08	L	0	0	0	10	5	20	25
S09	R	10	15	10	15	10	10	0
S10	R	15	10	10	5	15	25	20
S11	L	10	10	5	0	5	5	5
S12	L	20	15	10	0	5	10	10
S13	R	20	15	5	15	10	10	15
S14	L	30	15	10	10	30	10	0
S15	L	25	15	10	5	0	25	20
S16	L	15	10	0	10	25	25	10
S17	L	20	10	15	0	5	20	20
S18	R	5	5	0	5	0	15	0
S19	L	5	5	5	5	20	10	15
S20	R	5	5	0	0	5	15	10
S21	L	30	15	15	30	10	10	15
S22	R	15	15	10	20	-10	15	0
S23	R	10	5	5	10	5	10	0

of the notched noise to mask possible low-frequency combination bands [24, 25]. An additional 16th-order lowpass Butterworth filter with a cutoff frequency of $(1 - r) f_c$ was added to prevent this lowpass noise from appearing in the notch. The noise duration was 700 ms (including 30 ms cosine onset and offset ramps), and the probe tone duration was 500 ms (including 10 ms cosine onset and offset ramps), starting 100 ms after the noise onset.

2.3.2. Procedure

Detection thresholds for the probe tone were obtained using a two down, one up, two-interval, two-alternative forced-choice (2I-2AFC) paradigm to estimate the 70.7% point on the psychometric function [26]. The probe tone level was held constant at 63 dB SPL at 370 Hz, and at 44 dB SPL at 1394 Hz. These levels were chosen to match the mean spectrum levels of the two first formants of the six vowels used in the streaming test. At the beginning of the procedure, the probe tone and the masker had the same spectrum level. The masker level was then adjusted in accord with the response of the participant. The initial step size was 8 dB prior to the two first turnarounds, then 4 dB for 2 turnarounds, and finally 2 dB for eight turnarounds. These eight turnarounds were averaged to compute the threshold. The thresholds were measured for at least three notch ratios per participant, in random order, and always including 0.0. The other values were determined individually for each subject in order to avoid overexposure and saturation, and were always smaller than 0.2. Larger ratios are typically used in the literature, but with a fixed level probe, the level of the masker rapidly becomes a limitation for wider notches. For subjects S01 to S05, three measurements per f_c were performed, whereas at least six measurements were performed for subjects S06 to S23.

A fitting procedure was performed to derive auditory filter shapes from the data, using a symmetric roex(p) model [23, 27] without pedestal: $(1 + pg)\mathrm{e}^{-pg}$, with g the normalized distance from the center of the filter. The average spectrum of 4000 repetitions of the masker noise was used for integration under the filter shape. The fitting procedure took into account the Sennheiser HD250 Linear II transfer function, the middle-ear transfer function, and variations in filter bandwidth with center frequency. The equivalent-rectangular bandwidth of the auditory filter (ERB [23]) was then computed from the fitting, and expressed as the ratio to the ERB_N [28]. The ERB centered on 370 Hz and 1394 Hz are noted ERB_{370} and ERB_{1394} respectively. To retain only plausible auditory filter widths, values below $0.5\,\mathrm{ERB}_N$ or above $3.0\,\mathrm{ERB}_N$ were excluded from the analyses. The reliability of each ERB measurement was estimated by computing the ERB distribution as follows. For each measurement, 500 fittings were performed replacing the thresholds by random values drawn from normal distributions centered on the measured thresholds and having a standard deviation equal to the standard deviation of the eight last turnarounds. The standard deviation of the obtained ERB distribution was used as a reliability measure for each ERB estimate.

2.4. Streaming with vowels

The method used to evaluate streaming of a sequence of vowels was based on an order-naming task [10, 29]. In this task, sequences of vowels with alternating F_0 are presented and the subject is asked to report the vowels in the correct order. When segregation occurs, the perception of order is lost across the auditory streams, rendering the task difficult or impossible. This paradigm therefore provides an objective estimation of obligatory streaming [10]. The term “obligatory” is used here to indicate that the task reflects streaming that cannot be suppressed by the listener, as accurate performance is hindered by streaming.

2.4.1. Materials

The materials were built and used in another study [30] and consisted of recorded rather than synthesized vowels (as were used in [10]). The six French vowels /a e i ɔ u y/ were recorded at 24 bits and 48 kHz, using a Røde NT1 microphone, a Behringer Ultragain preamplifier, a Digigram VxPocket 440 soundcard and a PC. The speaker was instructed to pronounce all six vowels at the same pitch, and to reduce prosodic variations.

The F_0 and duration of each vowel was then manipulated using STRAIGHT [31]. Duration was set to 165 ms, including 10 ms raised cosine onset and offset ramps, which approximates the average syllable rate in French and English [32]. Average F_0 was adjusted to 100, 134, 179 and 240 Hz. Fundamental frequency variations related to intonation were constrained to 0.7 semitones from the average F_0, and formant positions were held constant across F_0s.

The vowels were then concatenated to form sequences. Each sequence contained one presentation of the six different vowels. The F_0 of the vowels alternated between two values $F_{0(1)}$ and $F_{0(2)}$. For all sequences, $F_{0(1)}$ was 100 Hz, and $F_{0(2)}$ was one of the following values: 100, 134, 179 or 240 Hz. Each sequence was created with two presentation rates: Slow at 1.2 vowel/s and Fast at 6 vowel/s. Slow sequences were created by inserting silence between the vowels, and were used to check vowel identification performance. Fast sequences were used to observe streaming. Finally, each sequence was repeated to form the final stimuli. Slow sequences were repeated four times, and Fast sequences were repeated 20 times, for overall durations of 20 s. For each possible arrangement of the 6 vowels a *perceptual distance* between formants of successive vowels was calculated from the formant frequencies expressed on a Bark scale (see [10] for details). The 40 arrangements of six vowels having the lowest *perceptual distance* were selected for inclusion, and these orders were assigned to F_0 conditions such that average *perceptual distance* was equivalent across conditions. This was performed to enhance the influence of F_0 differences across alternate vowels, and reduce the influence of streaming based on differences in formant structure [29]. Stimuli were generated with 16 bits and 44.1 kHz using MATLAB, and presented at 85 dB SPL.

2.4.2. Procedure

Training: The training began with a simple identification task on single vowels. Each vowel, at each F_0 (100, 134, 179 and 240 Hz), was presented twice in random order. Visual feedback was provided after each response. All subjects achieved more than 93% correct. The second step of training involved another form of vowel identification. In this step, vowels were presented in Slow sequences. In each of two blocks, 20 sequences were presented, 5 at each F_0. The procedure was the same as the test procedure described in the next paragraph, except that visual feedback was provided and that only Slow sequences were used. Performance in this second vowel identification task was 98% correct on average, ranging from 82% to 100%. Training lasted 17 min on average.

Streaming test: The streaming test was composed of two blocks of 40 sequences each. Half the sequences were in the Slow condition to check identification within the test, and half the sequences were in the Fast condition to examine streaming. Each sequence was presented to the subject during 20 s, but he/she was locked out from responding during the first 5 s, to allow the percept to stabilize. The subject then had to ‘Write the sequence in the correct order’ by selecting six times one vowel among the six possibilities using a mouse and computer graphic interface. The next sequence was presented after the subject had submitted their response or after the 20 s expired. The different conditions (vowel rate and F_0s) were presented in random order. The average duration of each block was about 12.5 min. For each subject, this procedure provides scores as a function of $F_{0(2)}$. The score is the percentage of sequences for which the subject successfully reported the six vowels in the correct order. Thus, high scores correspond to perception of a single integrated stream, while low scores correspond to segregation of the stimuli into two streams.

2.5. Speech-in-speech reception

2.5.1. Materials

The concurrent speech test consisted of target words presented simultaneously with a continuous masker. The masker was time reversed speech, used for its lack of semantic content. To create the masker, a male talker was digitally recorded (with the same settings and apparatus employed for the vowels in the previous section) reading a newspaper article for a duration of 5 min. Silences were deleted, and the root-mean-square (RMS) level was adjusted to be constant over 12 s Hann windows having 50% overlap. The resulting signal was then segmented into 45 s segments and downsampled to 24 kHz. Finally, each segment was processed with STRAIGHT to set the average F_0 to 100 Hz (min: 86 Hz, max: 116 Hz), and time reversed. The target speech was composed of lists of monosyllabic French words uttered by a different male speaker. The lists were extracted from the Vocales audio-CD as used in Hoen *et al.* [33]. The RMS level of all words was adjusted to 85 dB SPL. Words were arranged into 24 lists of 10 words each. The word lists were balanced in frequency of occurrence, phonological neighborhood, number of phonemes and duration. Using STRAIGHT again, the F_0 for each subset of six lists was set to 100, 134, 179, or 240 Hz.

As in the streaming test, formant positions were held constant. The level of the target words was fixed while the masker level was modified. In each F_0 condition, each of the six lists was combined with a masker to form a different target-to-masker ratio (SNR): −9, −6, −3, 0, 3 and 6 dB.

2.5.2. Procedure

The participants were asked to listen to the word lists and to repeat each word they heard. The lists (and F_0) were presented in random order. The pronounced words were written down by the experimenter and then converted into phonetic representations. A score – the proportion of correct phonemes in the words – was generated for each F_0 and SNR.

3. Results and discussion

3.1. Frequency selectivity

The fitting procedure produced ERB estimates smaller than 0.5 $\mathrm{ERB_N}$ or larger than 3.0 $\mathrm{ERB_N}$ at f_c = 370 Hz for subject S01, and at f_c = 1394 Hz for S20. These two estimates were therefore excluded from the analyses. The ERB of the auditory filters as a function of audiometric thresholds are displayed in Figure 1. The mean $\mathrm{ERB_{370}}$ was 87.3 Hz (1.3 times the normal ERB), and the mean $\mathrm{ERB_{1394}}$ was 293 Hz (1.7 times the normal ERB). These mean values are relatively consistent with those reported by Moore [34], although they are somewhat larger, likely due to the fact the actual spectrum of the masker was used in the fitting procedure rather than a simplified version with infinite slopes. The present results also show a larger range of ERB values across subjects than that reported by Moore [34]. The current measurement involved relatively small r notch-ratios compared to those usually employed [24] and this could have emphasized the error in the fitting procedure and then in the ERB estimate. This is illustrated by the error bars in Figure 1. Additional statistical analyses were therefore performed using the difference between the raw thresholds measured at $r = 0.0$ and at $r = 0.1$ as a measure of frequency selectivity. Because the choice of the frequency selectivity measurement did not affect the effects or correlations, the ERB estimate was retained, despite its potential limitations, because it is a more common representation of frequency selectivity. Finally, the p-values in the correlation analyses have been corrected to take into account the estimated ERB variability.

3.2. Streaming with vowels

The results of the streaming test are plotted in Figure 2. The average identification score in the control (Slow) condition is over 95% correct for all values of $F_{0(2)}$, and only subject S01 had scores lower than 90%. As described in [10], the scores in the Fast condition reflect segregation. The scores decreased from 57% for $F_{0(2)}$ = 100 Hz, to 13% for $F_{0(2)}$ = 240 Hz. A first-order one-way repeated measure ANOVA with $F_{0(2)}$ as a repeated parameter revealed a significant effect of $F_{0(2)}$ [$F(1,22)$ = 62.36,

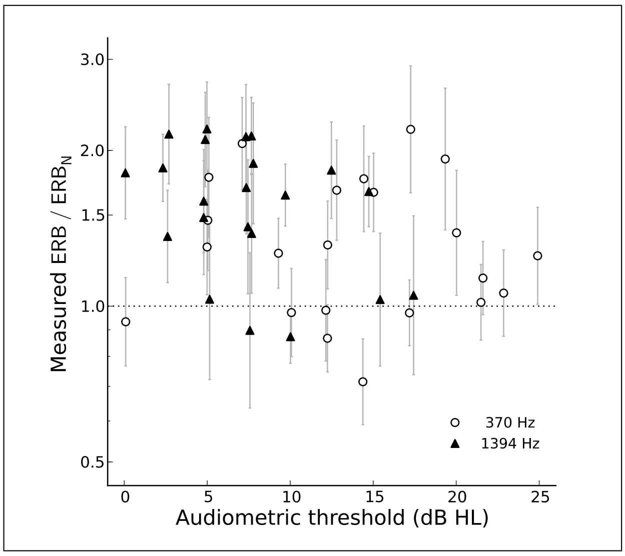

Figure 1. Values of the ERB of the auditory filter for each subject plotted as a function of absolute threshold (dB HL) interpolated at the test frequency. The ERB values are plotted relative to the ERB_N [23, 28]. For test frequency 370 Hz (open circles), the fitting procedure succeeded for 22 of the 23 participants. For test frequency 1394 Hz (filled triangles), the fitting procedure also succeeded for 22 of the 23 participants. The error bars display the standard deviation of ERB ratio estimates (see text for details).

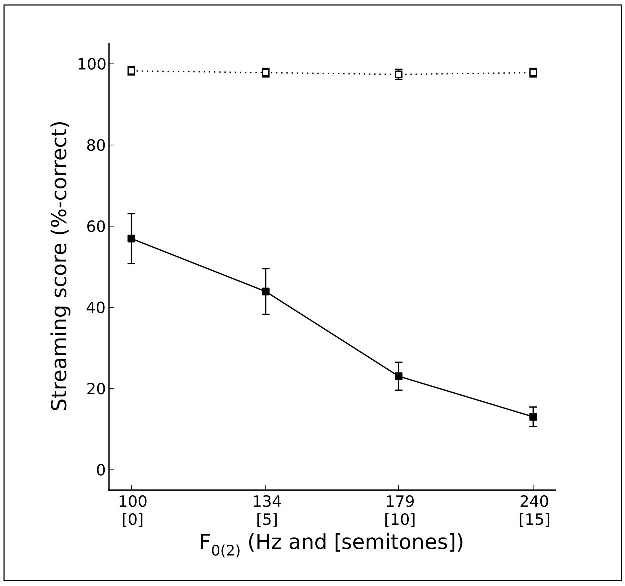

Figure 2. Streaming scores, in percent, averaged across participants as a function of $F_{0(2)}$ (in Hertz and in semitones re 100 Hz). The score is the percentage of sequences for which the subject reported the six vowels in the correct order. Streaming disrupts this ability, therefore lower scores reflect additional streaming. The scores for the Fast condition (streaming) are plotted with the solid line and filled squares. The scores for the Slow condition (vowel identification control condition) are plotted with the dashed line and open squares. The error bars represent the standard error across participants.

$p < .001$]. This reflects the effect of the F_0 difference on streaming: the greater the F_0 difference, the more segregation occurs. No difference was found between the two testing blocks [$F(1, 22) = 0.15$, $p = 0.70$], indicating that no substantial training took place during the session. The scores are consistent with those observed in [10]. In particular, the score at matched F_0 is similar to that observed in naïve NH listeners with synthetic vowels having a duration of 175 ms. It has been argued [10] that the score in this particular condition reflects streaming induced by formant structure, as described by Dorman *et al.* [29]. Note that the decrease in performance with increasing $F_{0(2)}$ is unlikely to be due to a reduction in vowel intelligibility since the identification scores (as measured in the Slow condition) were constant and high for all values of $F_{0(2)}$ [$F(1, 22) = 0.23$, $p = 0.64$]. Moreover, in another study Gaudrain *et al.* [30] used an $F_{0(1)}$ of 240 Hz and decreased $F_{0(2)}$ from 240 Hz to 100 Hz, and observed the same pattern of results.

In the following correlation analyses involving ERB, two measures of streaming are used: the average score in the streaming task noted ⟨Streaming⟩, and the ΔF_0 benefit in the streaming task noted DStreaming. The former is the mean score for a listener across the four values of $F_{0(2)}$. The latter is defined for each subject as the difference between the highest and the lowest score across F_0s. So the ΔF_0 benefit is the maximal decrease (as a positive value) in performance induced by changing ΔF_0.

3.3. Speech-in-speech reception

The results of the speech reception test are displayed in Figure 3. The scores are the percentage of phonemes correctly recalled, at each SNR and F_0. Without any F_0 difference between the target and masker, the mean score across SNRs was 67% correct. For F_0 differences greater than 5 semitones (134 Hz), the mean score was above 80%. A first-order two-way repeated measure ANOVA using SNR and F_0 as repeated parameters indicated that the effect of the F_0 difference was significant [$F(1, 22) = 162.53$, $p < .001$], as was the effect of SNR [$F(1, 22) = 288.34$, $p < .001$] and the interaction [$F(1, 22) = 16.26$, $p < .001$].

The present results are very similar to those obtained by Brokx and Nooteboom [3]. These authors found an increase in identification scores from about 40% to 60% for a 0 to 3 semitone difference. The scores observed in the current experiment are slightly greater, even when comparing percentage of words correct (50% to 64%) rather than scores based on phonemes. This is probably due to the fact that SNR ranged from −9 to 6 dB in the current experiment while it ranged from −15 to 0 dB in [3]. The benefit of F_0 difference for the identification of the target words depends on the SNR as revealed by the significant interaction. The tendency is that the benefit becomes smaller as the SNR increases. Also worth noticing is the fact that most of the benefit is achieved for F_0 differences smaller than five semitones (134 Hz), while adding another five-semitone difference increases performance only slightly and only for positive SNRs. This probably reflects the importance of simultaneous segregation, especially at lower SNRs, and is a reminder that both sequential and simultaneous segregation mechanisms are involved in speech-in-speech perception.

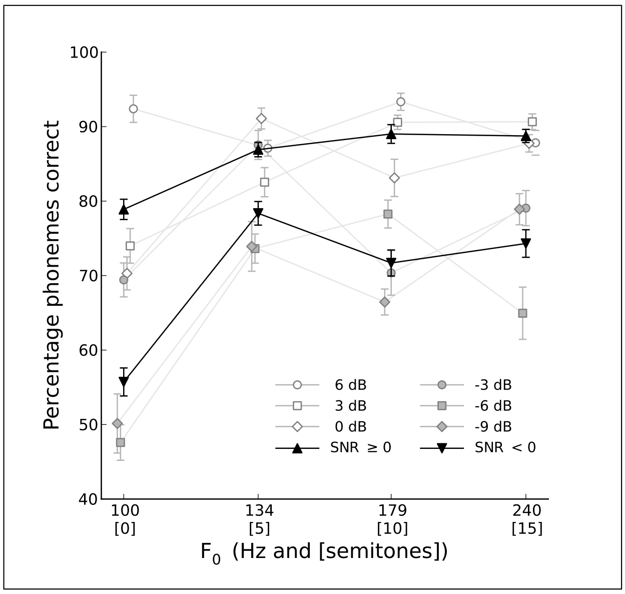

Figure 3. Speech identification scores as the proportion of correctly reported phonemes, averaged across participants, as a function of the F_0 of the target. Each light grey line represents a different SNR, while the two black lines represent the data averaged over positive (upward black triangles) and negative (downward black triangles) SNRs. The lower axis represents the F_0 of the target, the masker F_0 being always 100 Hz. The F_0 difference in semitones is provided between brackets. The error bars are the inter-individual standard error.

In the following correlation analyses involving ERB, the average speech-in-speech perception score (⟨Speech⟩) differs from the ΔF_0 benefit (ΔSpeech) as for the streaming task. The former is the mean score for a listener, averaged across $F_{0(2)}$s and SNRs. The latter is defined for each subject as the average across SNRs of the difference between the highest and the lowest mean score across F_0s.

4. Correlations and general discussion

4.1. Speech-in-speech perception and streaming

An F_0 difference of 5 to 15 semitones yielded an improvement in speech-in-speech perception scores, and increased the amount of streaming (*i.e.* was detrimental to streaming scores). If streaming is a mechanism underlying speech-in-speech segregation, a relationship between scores in these two tasks may be expected. The scores obtained in the speech task are plotted against those obtained in the streaming task for each listener and F_0 in Figure 4. A within-subject regression analysis revealed a highly significant correlation between the effect of F_0 on both tasks [$r = -0.72$, $F(1, 22) = 23.43$, $p < 0.001$]. This correlation illustrates the fact that listeners who benefit most from an F_0 difference between target and masker

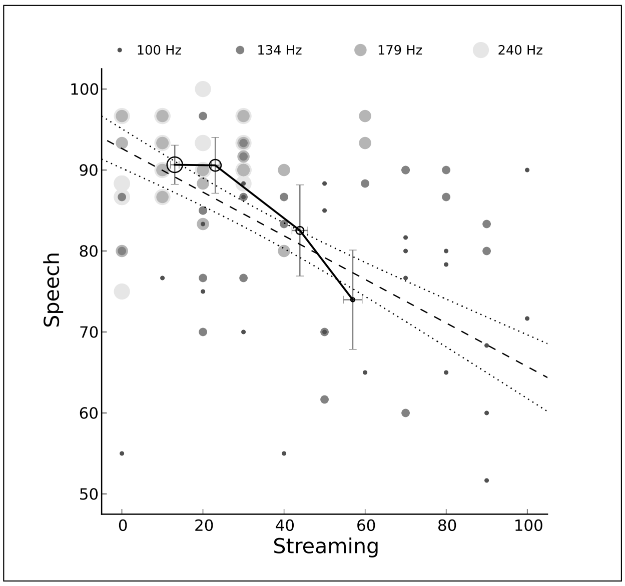

Figure 4. Speech scores against Streaming scores. The filled circles represent individual data for each F_0, their diameter and color is coding the specific value as shown in the figure legend. The black solid line represents the scores averaged across participants for each F_0 (error bars show inter-individual standard errors). The size of the open black circles codes the F_0 value. The dashed line shows the average within-participant correlation between the two variables, and the dotted lines shows the associated 95% confidence interval.

in speech-in-speech perception are also the ones who experienced the most segregation in the streaming task. This result is in accord with those of Mackersie *et al.* [11], who also found a relationship between fusion threshold for pure tones and concurrent sentence perception. Similarly, Hong and Turner [13] found a relationship in cochlear implant users between obligatory streaming of pure tones and speech-in-noise perception. The current results confirm that a similar relationship can be observed in close to normal hearing listeners exposed to speech stimuli. Furthermore, since the relationship concerns the effect of F_0, the current results suggest that the F_0-based streaming mechanism is involved in speech-in-speech perception. Since the benefit generated by small ΔF_0s in the speech-in-speech experiment seemed to reflect the fact that simultaneous segregation was involved, these results also suggest that the two segregation mechanisms may interact synergistically. Further investigations would be required to clarify this interaction.

4.2. Frequency selectivity and speech-in-speech perception

The effects of F_0 and frequency selectivity on speech-in-masker perception have been studied separately using various paradigms that have yielded various results. Festen and Plomp [15] found a correlation between speech-in-noise perception and the logarithm of the auditory bandwidth estimated using a comb-filtered noise masker and a probe tone. Similarly, Glasberg and Moore [16] measured SRT in quiet and in speech-shaped noise, and found a correlation between the SRT in noise and some measures related to the perception of frequency: tonal frequency difference limens, fundamental frequency difference limens, and the ERB. In these two studies, it was argued that speech-in-quiet perception relies largely on the audiometric threshold, while speech-in-noise perception depends on supra-threshold abilities such as spectral resolution. Mackersie *et al.* [11] used simultaneous sentences having F_0s of 115 and 240 Hz. In contrast to the previous literature [15, 16] these authors did not find any significant correlation between the slopes of the notched-noise masking function (a representation of frequency selectivity) and the percentage of words correct in target sentences. Mackersie *et al.* [11] argued that simultaneous sentences contain more contextual evidence and acoustic variability than the steady-state noise maskers used in previous studies, for which peripheral masking would have been enhanced.

To isolate the effect of frequency selectivity, many researchers first partialed out the effect of audiometric threshold. In the current experiment, the hearing losses are mild and so audibility should not have a substantial influence. Indeed, no correlation was found between mean audiometric threshold and mean speech perception scores [$r = 0.12$, $t(21) = 0.58$, $p = 0.57$]. Hence, the variations in audiometric threshold should not influence the ERB effect analysis.

In the current study, speech-in-speech identification was evaluated as a function of the ERB at 370 and 1394 Hz, as plotted in Figure 5. No significant correlation was observed between the average speech-in-speech perception score ⟨Speech⟩ and the ERB_{370} [$r = -0.23$, $t(20) = -1.06$, $p = 0.30$, $p_{corr} = 0.47$[1]], but a significant correlation was observed with the ERB_{1394} [$r = -0.49$, $t(20) = -2.51$, $p = 0.021$, $p_{corr} = 0.043$]. A similar analysis on the F_0 benefit scores ΔSpeech revealed no significant effect of the ERB value at 370 Hz [$r = 0.07$, $t(20) = 0.33$, $p = 0.75$, $p_{corr} = 0.92$] or at 1394 Hz [$r = 0.20$, $t(20) = 0.93$, $p = 0.37$, $p_{corr} = 0.57$].

Frequency selectivity in the region of the second formant correlated with overall intelligibility. A first explanation for why a significant correlation was found only at 1394 Hz might be found in the amount of information transmitted to the auditory pathway in each frequency channel. Stilp and Kluender [35] have borrowed the concept of entropy from information theory and showed that entropy estimated at the cochlear level can be used as a reliable measure of the amount of phonological information transmitted. They implemented their entropy measure using the Euclidian distance between successive excitation patterns as evaluated by computational auditory fil-

[1] p_{corr} is the p-value corrected for ERB variability. Five thousand random ERB data sets centered on the measured ERB and reflecting the estimated ERB variability were generated. For each data set, the correlation coefficients with ⟨Speech⟩ and ΔSpeech were calculated, thus building distributions for these statistics. The two-tailed probability of having a correlation coefficient different from zero was drawn directly from the distribution and used to adjust the original p-value.

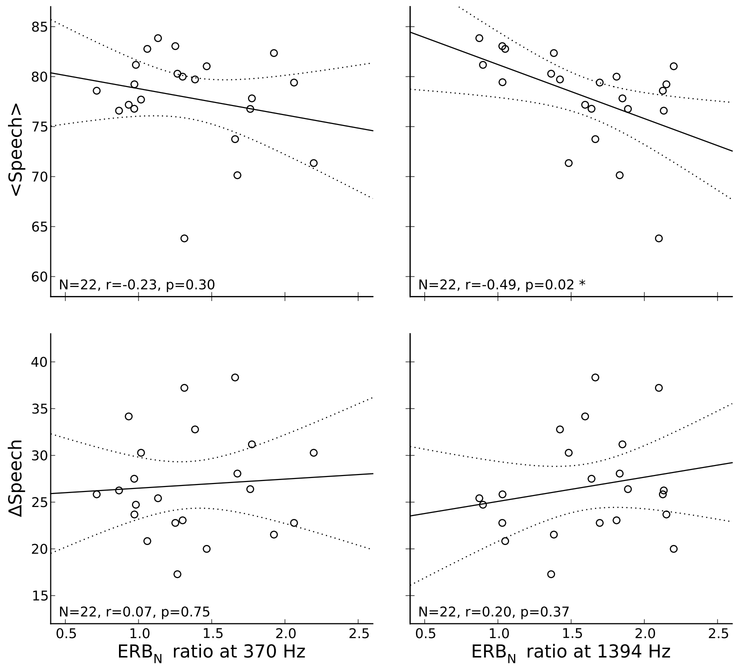

Figure 5. Speech-in-speech perception score as function of the ERB. In all panels, the circles represent individual data, the solid line shows the correlation, and the dotted curves show the 95% confidence interval of the correlation The ERB is displayed as the ratio relative to the ERB_N. Upper panels: ⟨Speech⟩, i.e. average speech-in-speech perception score across target F_0 and SNR, for each listener. Lower panels: ΔSpeech, i.e. F_0 speech benefit. Left panels: data for 22 listeners for the ERB centered on 370 Hz. Right panels: data for 22 listeners for the ERB centered on 1394 Hz.

terbanks. We adapted this measure by using a single auditory filter centered on a region of interest to derive a frequency-specific entropy and applied it to the speech material used in the present experiment. Our measurements showed that the entropy in the auditory filter centered on 1394 Hz was 30 to 60% larger than that at 370 Hz. This indicates that the frequency region around 1394 Hz carried more information than that around 370 Hz. However it cannot be excluded that the absence of correlation with ERB_{370} was due to the reduced reliability of this measure. Indeed, while a notch of $0.2f_c$ (the largest notch width used) represents $1.6\,\mathrm{ERB}_N$ at 1394 Hz, it only represents $1.1\,\mathrm{ERB}_N$ at 370 Hz. The reduction in reliability due to the small notch sizes used in the current experiment is therefore probably more pronounced at 370 Hz than at 1394 Hz.

The effect of frequency selectivity at 1394 Hz on overall performance is consistent with what has been observed for speech-in-noise perception [15, 16]. But this result contrasts with that of Mackersie *et al.* [11] who found no relationship between simultaneous sentence perception and frequency selectivity. Mackersie *et al.* suggested that the absence of correlation could be due to the fact that their simultaneous sentence material offered more contextual information than when a noise masker is used, thus reducing the importance of peripheral masking. In the present study, time-reversed speech was used, which, like noise maskers, offered little or no context. This may potentially emphasize the role of peripheral masking provided by the fluctuating masker.

Although the overall level of identification appeared to be related to frequency selectivity, the segregation benefit that originates from an F_0 difference was not found to be related to frequency selectivity. Instead this benefit might be more directly related to F_0 discrimination abilities, despite that previous studies observed no clear relationship [4]. This point is discussed in the next section, along with the streaming results.

4.3. Frequency selectivity and streaming with vowels

The peripheral channeling theory [36] hypothesized a relationship between frequency resolution and streaming, and a few studies have attempted to observe this effect. Rose and Moore [12] measured the fission boundary for pure tones in listeners with NH and in listeners with unilateral cochlear hearing loss. For the NH listeners, they found that the frequency difference at the fission boundary (where the percept changes from two streams to one stream) was constant across frequency when expressed in ERB_Ns. However, they observed no clear fission boundary difference across ears of the unilaterally-impaired listeners, suggesting that fission boundary might not be directly related to frequency selectivity. More recently, these authors compared the fission boundary to frequency difference limens in NH and HI listeners [37]. They observed, for the NH listeners, that the fission boundary was fairly constant at eight times larger than the frequency difference limen, in the 250–2000 Hz region. They did not find such a clear relationship in HI listeners, however enlarged frequency difference limens may have contributed to elevated fission boundaries, and other factors were also likely involved. Grimault *et al.* [19] also observed streaming in NH and HI listeners, using resolved and unresolved complex tones. They found that complex tones that were unresolved for both NH and HI listeners yielded similar streaming performance, while complex tones that were resolved for NH but not for HI induced more streaming in NH than in HI. Gaudrain *et al.* [10] used a spectral smearing algorithm to simulate auditory filter broadening [17] in a streaming task

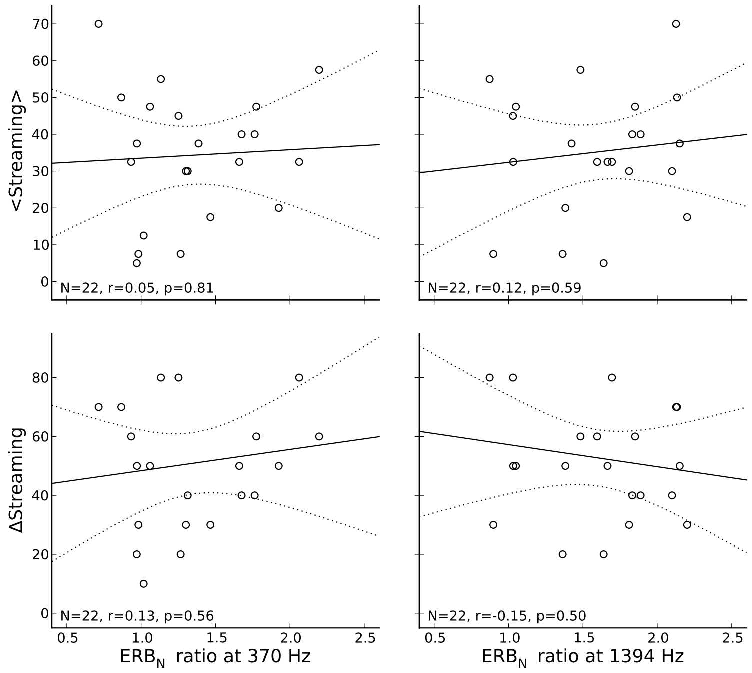

Figure 6. Same as Figure 5 with ⟨Streaming⟩ (upper panels) and ΔStreaming (lower panels).

that involved synthetic vowels. They observed that spectral smearing three times normal hindered F_0-based obligatory streaming.

In the current study, streaming was compared to the measured ERBs (Figure 6). The average score in the streaming task (⟨Streaming⟩), was not related to ERB_{370} [$r = 0.05$, $t(20) = 0.25$, $p = 0.81$, $p_{corr} = 0.97$] or ERB_{1394} [$r = 0.12$, $t(20) = 0.55$, $p = 0.59$, $p_{corr} = 0.82$]. A similar analysis on ΔStreaming, the streaming score improvement due to the F_0 difference, revealed no significant relationship at ERB_{370} [$r = 0.13$, $t(20) = 0.59$, $p = 0.56$, $p_{corr} = 0.77$] or at ERB_{1394} [$r = -0.15$, $t(20) = -0.69$, $p = 0.50$, $p_{corr} = 0.73$].

Neither the ΔF_0 benefits nor the average streaming scores were correlated with either of the ERB measures. This result is consistent with those obtained using pure tones by Rose and Moore [12] and Mackersie *et al.* [11]. However, they are in contrast with Gaudrain *et al.* [10], where simulated broad auditory filters were found to significantly hinder streaming in a task similar to that employed here. More specifically, Gaudrain *et al.* reported that spectral smearing improved both the mean scores and the ΔF_0 benefit in streaming (reflecting less streaming). The authors argued that spectral smearing hindered streaming based on F_0 difference as well as streaming based on formant structure. Overall performance in the streaming task depends on F_0-based and formant structure-based streaming, and also reflects the ability of listeners to perform the order-naming task. In contrast, the ΔF_0 benefit reflects solely the effect of F_0-based streaming.

The fact that the ΔF_0 benefit was not correlated with frequency selectivity in the current study may suggest that F_0-based streaming relies on another psychoacoustic factor. As found by Rose and Moore [37] with pure tones, F_0-based streaming probably depends on F_0 discrimination performance, which in turn has been found to be only weakly correlated with frequency selectivity [14]. The same lack of relationship was found between frequency selectivity and the concurrent speech perception task. Since the ΔF_0 effect was correlated between the streaming and speech tasks, it seems reasonable to postulate that the two are driven by a common mechanism related to pitch perception.

The discrimination of F_0 in complex tones, or pitch perception in a more general sense, involves the perception of temporal cues: temporal envelope periodicity and fine structure [38, 39]. The ability of subjects to benefit from these cues may not be highly correlated with frequency selectivity, especially when frequency selectivity is close to normal. The spectral smearing algorithm of Baer and Moore [17] used by Gaudrain *et al.* [10] to simulate broadened auditory filters mimics the spectral aspect of frequency selectivity impairment, but the time windowing also markedly alters temporal fine structure. Hence the effect of spectral smearing observed in Gaudrain *et al.* [10] could potentially have been caused by both the degradation of frequency selectivity and the degradation of temporal fine structure. Furthermore, it has been suggested that HI listeners have only limited access to temporal fine structure cues (e.g., Lorenzi *et al.* [40]). These authors also demonstrated that the performance of these subjects in concurrent speech perception was correlated with their ability to use temporal fine structure. In the current study, the ability to use temporal fine structure probably varied across participants, perhaps independently of frequency selectivity. Further investigation is required to assess this hypothesis.

5. Conclusions

1. The ΔF_0 benefit in streaming and the ΔF_0 benefit in speech-in-speech perception were correlated, suggesting that they are driven by a similar underlying factor. However, neither benefit was correlated with the ERB suggesting that, at least in the current experiment, frequency selectivity was not the underlying common factor. The source of the common variability between the two measures thus remains to be explained.
2. Average streaming scores did not correlate with frequency selectivity, while average speech-in-speech performance did only for the auditory filter centered on 1394 Hz. Although average streaming scores and average speech-in-speech performance are both directly related to speech intelligibility, frequency selectivity seemed to only affect speech-in-speech. This thus indicates that the significance of the relationship between speech intelligibility and frequency selectivity may vary depending on the material and task used.
3. The nonsignificant correlations involving ERB measurements should be interpreted with caution because (i) the reduced notch widths employed provide limited measurement reliability, and (ii) the participants had close to normal hearing which reduced the range of ERBs.

Acknowledgements

The authors wish to thank Claire Grataloup and François Pellegrino for providing word list materials and support in building the speech-in-speech experiment, Joshua G. W. Bernstein for providing a MATLAB implementation of the auditory filter function fitting procedure, Brian C. J. Moore for advice on notched-noise data interpretation, and the Associate Editor and two anonymous reviewers for particularly helpful comments on a previous version of the manuscript. This work was supported in part by NIH grant DC08594, by a doctoral grant from the Région Rhône-Alpes (France), by grant JC6007 from the Ministère de l'Enseignement supérieur et de la Recherche (France), and by grant ANR-08-BLAN-0167-01 (project Multistap) from the Agence Nationale de la Recherche (France).

References

[1] A. S. Bregman: Auditory scene analysis: The perceptual organization of sound. The MIT Press, Cambridge, MA, 1990.

[2] E. C. Cherry: Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. **25** (1953) 975–979.

[3] J. P. L. Brokx, S. G. Nooteboom: Intonation and the perceptual separation of simultaneous voices. J. Phonetics **10** (1982) 23–36.

[4] V. Summers, M. R. Leek: FO processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss. J. Speech Lang. Hear. Res. **41** (1998) 1294–1306.

[5] J. Bird, C. J. Darwin: Effects of a difference in fundamental frequency in separating two sentences. – In: Psychophysical and physiological advances in hearing. A. R. Palmer, A. Q. Summerfield, R. Meddis (eds.). Whurr, London, 1998.

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

[7] P. F. Assmann, Q. Summerfield: Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J. Acoust. Soc. Am. **88** (1990) 680–697.

[8] R. Meddis, M. J. Hewitt: Modeling the identification of concurrent vowels with different fundamental frequencies. J. Acoust. Soc. Am. **91** (1992) 233–245.

[9] L. P. A. S. van Noorden: Temporal coherence in the perception of tones sequences. Eindhoven University of Technology, The Netherlands, 1975.

[10] E. Gaudrain, N. Grimault, E. W. Healy, J.-C. Béra: Effect of spectral smearing on the perceptual segregation of vowel sequences. Hear. Res. **231** (2007) 32–41.

[11] C. L. Mackersie, T. L. Prida, D. Stiles: The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss. J. Speech Lang. Hear. Res. **44** (2001) 19–28.

[12] M. M. Rose, B. C. J. Moore: Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners. J. Acoust. Soc. Am. **102** (1997) 1768–1778.

[13] R. S. Hong, C. W. Turner: Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients. J. Acoust. Soc. Am. **120** (2006) 360–374.

[14] B. C. J. Moore, R. W. Peters: Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity. J. Acoust. Soc. Am. **91** (1992) 2881–2893.

[15] J. M. Festen, R. Plomp: Relations between auditory functions in impaired hearing. J. Acoust. Soc. Am. **73** (1983) 652–662.

[16] B. R. Glasberg, B. C. J. Moore: Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. Scand. Audiol. Suppl. **32** (1989) 1–25.

[17] T. Baer, B. C. J. Moore: Effects of spectral smearing on the intelligibility of sentences in noise. J. Acoust. Soc. Am. **94** (1993) 1229–1241.

[18] T. Baer, B. C. J. Moore: Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. J. Acoust. Soc. Am. **95** (1994) 2277–2280.

[19] N. Grimault, C. Micheyl, R. P. Carlyon, P. Arthaud, L. Collet: Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing. Br. J. Audiol. **35** (2001) 173–182.

[20] American National Standards Institute: Specifications for audiometers S3.6-2004. American National Standards Institute, New York, USA, 2004.

[21] American National Standards Institute: Methods for coupler calibration of earphones S3.7-1995 (r2003). American National Standards Institute, New York, USA, 1995.

[22] R. D. Patterson: Auditory filter shapes derived with noise stimuli. J. Acoust. Soc. Am. **59** (1976) 640–654.

[23] B. R. Glasberg, B. C. Moore: Derivation of auditory filter shapes from notched-noise data. Hear. Res. **47** (1990) 103–138.

[24] J. G. W. Bernstein, A. J. Oxenham: The relationship between frequency selectivity and pitch discrimination: Sen-

sorineural hearing loss. J. Acoust. Soc. Am. **120** (2006) 3929–3945.

[25] D. D. Greenwood: Masking by combination bands: Estimation of the levels of the combination bands $(n+1)f_1 - nf_h$. J. Acoust. Soc. Am. **52** (1972) 1144–1154.

[26] H. Levitt: Transformed up-down methods in psychoacoustics. J. Acoust. Soc. Am. **49** (1971) 467–477.

[27] R. D. Patterson, I. Nimmo-Smith, D. L. Weber, R. Milroy: The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. J. Acoust. Soc. Am. **72** (1982) 1788–1803.

[28] B. C. J. Moore: An introduction to the psychology of hearing. Fifth edition. Academic Press, 2003.

[29] M. F. Dorman, J. E. Cutting, L. J. Raphael: Perception of temporal order in vowel sequences with and without formant transitions. J. Exp. Psychol. Human **104** (1975) 147–153.

[30] E. Gaudrain, N. Grimault, E. W. Healy, J.-C. Béra: Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation. J. Acoust. Soc. Am. **124** (2008) 3076–3087.

[31] H. Kawahara, T. Irino: Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. – In: Speech separation by humans and machines. P. L. Divenyi (ed.). Kluwer Academic, Massachusetts, 2004, 167–180.

[32] A. D. Patel, J. R. Iversen, J. C. Rosenberg: Comparing the rhythm and melody of speech and music: the case of British English and French. J. Acoust. Soc. Am. **119** (2006) 3034–3047.

[33] M. Hoen, F. Meunier, C.-L. Grataloup, F. Pellegrino, N. Grimault, F. Perrin, X. Perrot, L. Collet: Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. Speech Comm. **49** (2007) 905–916.

[34] B. C. J. Moore: Cochlear hearing loss: physiological, psychological and technical issues. Wiley-Interscience, 2007.

[35] C. E. Stilp, K. R. Kluender: Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. P. Natl. Acad. Sci. USA **107** (2010) 12387–12392.

[36] W. M. Hartmann, D. Johnson: Stream segregation and peripheral channeling. Music Percept. **9** (1991) 155–183.

[37] M. M. Rose, B. C. J. Moore: The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects. Hear. Res **204** (2005) 16–28.

[38] G. A. Moore, B. C. J. Moore: Perception of the low pitch of frequency-shifted complexes. J. Acoust. Soc. Am. **113** (2003) 977–985.

[39] K. Hopkins, B. C. J. Moore: Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information. J. Acoust. Soc. Am. **122** (2007) 1055–1068.

[40] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, B. C. J. Moore: Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc. Natl. Acad. Sci. U.S.A. **103** (2006) 18866–18869.