

Influence of Competing Multi-Talker Babble on Frequency-Importance Functions for Speech Measured Using a Correlational Approach

Gaëtan Gilbert*, Christophe Micheyl†
UMR CNRS 5020, Lyon I University, Lyon 69366, France. gaetan@ihr.gla.ac.uk

Summary

In this study, a correlational approach was used to estimate the relative importance of five different frequency bands for the identification of speech in two listening conditions: In the first, the nonsense vowel-consonant-vowel (VCV) signals that the listener had to identify were presented in the absence of other potentially interfering speech signals. In the other condition, the target VCVs were presented in a multi-talker babble background. The signal-to-babble ratio (SBR) was fixed at 5 dB. Fifteen young normal-hearing listeners were tested. Each listener was successively presented with 1500 stimuli in each condition. On each presentation, a randomly selected VCV signal was added to noise in five contiguous frequency bands within a range from about 100 to 7750 Hz. The signal-to-noise ratio (SNR) was varied independently in the different bands over a 24 dB range, the mid-point of which was adjusted so that listeners achieved between 60 and 70% correct in both test conditions. The importance of each frequency band was estimated by computing the point bi-serial correlation coefficient between the successive SNRs in that band and the corresponding binary identification scores (correct/incorrect) across the 1500 trials. The shapes of the importance functions measured in the two conditions were found to be significantly different: the importance of the middle-to-high frequency band was significantly lower in the presence of babble, while the importance of the two lowest bands was significantly higher. Possible reasons for this specific finding and more general considerations regarding the application of the correlational method are discussed.

PACS no. 43.71.An, 43.71.Bp, 43.71.Es, 43.71.Gv

1. Introduction

Although speech perception has been the object of intense research over the past fifty years or so, how our ears and brains achieve high levels of speech recognition in adverse listening situations remains largely unknown. Speech is a highly complex signal, which contains a wealth of information scattered in both the temporal and the spectral domains. As a result, it is often a challenging task to determine which of the multiple cues present in the signal are effectively utilized by the central nervous system in order to achieve correct recognition of speech in quiet, let alone in the presence of noise or competing speech signals.

A few years ago, Doherty and Turner [1] and Turner *et al.* [2] devised a method for estimating the relative importance – or weights – of different frequency bands for speech recognition in a given individual. The method involves the addition of random amounts of noise in the

different frequency bands of the speech signals that the listener must identify. On each stimulus presentation, the signal-to-noise ratio (SNR) in each band is varied independently of that in the other bands. Following each stimulus presentation, the listener's score (correct or incorrect) is recorded. At the end of the experiment, the point bi-serial correlation coefficients between the binary identification scores (correct/incorrect) measured across trials and the corresponding SNRs in the different bands are computed. The correlation coefficients thus obtained provide a measure of the strength of the (linear) relation between the SNRs in the different bands and the speech identification scores. In that sense, the coefficients may be regarded as a measure of the importance of the different bands for speech identification, under the conditions and in the listener(s) tested. In principle, the larger the absolute value of the correlation coefficient between SNR and scores for a given band, the larger the contribution of that band to identification. Negative correlations, when they are observed, suggest that the presence of energy or information in the considered band is detrimental, rather than beneficial, to correct identification of the speech material.

The correlational approach proposed by Turner and colleagues may provide an interesting alternative to the filtering approach traditionally used to measure frequency-

Received 4 February 2004,
accepted 29 September 2004.

* Currently at the MRC Institute for Hearing Research, Glasgow Royal Infirmary, 16 Alexandra Parade, G31 2ER Glasgow

† Currently at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA, USA

importance functions in the context of the articulation or speech-intelligibility index [3, 4, 5, 6, 7]. The identification of low-, high-, or band-pass filtered speech may involve different listening strategies than those involved in broadband testing conditions, with energy present in all of the bands simultaneously. For instance, it could be the case that when two bands are presented together, information in one of the bands may hamper, or on the contrary facilitate, the processing of information in the other band. There is now ample empirical evidence for the existence of such negative or positive interactions between nearby or remote frequency bands in speech perception (e.g. [8, 9]). It has been pointed out that the AI or SII fails to account for such interactions [10, 11]. The correlational approach, involving the presentation of all bands simultaneously, goes some way toward taking into account simultaneous influences between bands when measuring importance functions.

An important question is whether and how the relative importance of different frequency bands for speech identification varies depending upon listening conditions. Speech recognition in different types of backgrounds is an important topic, which has been the object of numerous studies already (e.g. [12, 13]). Shifts in speech reception thresholds (SRTs) for different types of maskers, such as steady or fluctuating noise and concurrent speech, have been documented. However, the mechanisms behind these global effects remain uncertain. By providing insight into how a given interferer affects the contribution of different frequency bands to speech identification, the correlational method may help to better understand the perceptual mechanisms behind the interference. In the present study, frequency-importance functions for the identification of nonsense syllables were measured with the correlational method for normal-hearing listeners in two different listening conditions: in one condition, the speech signals that the listeners had to identify were presented in the absence of other potentially interfering speech signals; in the other condition, the target speech signals were presented in the presence of a competing multi-talker babble background. The main objective of the study was to test whether and how the introduction of the competing babble would alter the shape of the frequency-importance functions, i.e., the relative importance of different frequency bands.

2. Methods

2.1. Subjects

Fifteen subjects (aged between 19–27 years with a mean of 23.9 years) with normal hearing (pure-tone thresholds ≤ 20 dB HL at octave frequencies between 250 and 8000 Hz) took part in the study.

2.2. Stimuli and procedure

The basic speech stimuli used in this study consisted of VCV syllables, where V represents a vowel (a, i or u) – the same vowel was used in initial and final positions –, and

C represents a consonant, chosen (randomly) among 17 different consonants of the French language. Since most of the resulting 51 VCV combinations did not correspond to existing words in the French language, the corpus used in this study may be described as comprised of nonsense syllables. Each VCV combination was uttered four times by each of four different French-native talkers (two male and two female), leading to a total of 816 possible signals. The signals were digitally bandpass-filtered between 100 and 7750 Hz (4th-order IIR Butterworth filter). The amplitude of the VCV signals was normalized in such a way that the highest value of the root-mean-square (RMS) amplitude (computed within sliding rectangular windows of 92-ms duration each with 75% overlap between consecutive slices) was constant across signals.

On each presentation, one of the 816 VCV signals was selected. The signal was added to five rectangular bands of noise, defined by low and high corner frequencies of approximately 100–250 Hz (band 1), 250–750 Hz (band 2), 750–1750 Hz (band 3), 1750–3750 Hz (band 4), and 3750–7750 Hz (band 5), yielding bandwidths of approximately 150, 500, 1000, 2000, and 4000 Hz¹. These bands were selected arbitrarily, having in mind applications of the correlational approach to the evaluation of frequency-band-importance functions in the context of multi-band auditory prostheses. In that context, the method should be flexible enough to accommodate frequency bands whose cutoff frequencies and bandwidths are pre-defined based on prior settings or functional constraints, and cannot be modified at will by the experimenter. It is not uncommon for the widths of frequency bands in multi-band prostheses not to be constant on a linear frequency scale but, instead, to increase with center frequency. While it would certainly be interesting to study how importance functions measured using the correlational method are affected by how the bands are defined, due to the time consuming nature of the approach, we did not do it here. Most important, here, is that the exact same bands were used in the two testing conditions in which the measured importance functions were to be compared.

The RMS amplitude of each noise band was adjusted relative to the RMS amplitude of the target speech signal inside the considered frequency band; both RMS values were computed over the energetic part of the signal (silent parts at the beginning and at the end of the signal were deleted). On each stimulus presentation, the SNR in each band was selected randomly among 13 equally likely values, spanning a 24-dB range, from -12 to +12 dB (in 2-dB steps) around a pre-defined mid-point SNR. This mid-point SNR, which was equal across bands, was adjusted for each listener to yield an error rate of approximately 35%. Like in [1] and [2], this was achieved by first having the listener perform 200 trials at an SNR (the same across

¹ For technical reasons related to the sampling frequency and the number of bins in the FFT, the actual corner frequencies of the bands were slightly different from these indicated in the text. The exact corner frequencies used were: 97–248 Hz (band 1), 248–741 Hz (band 2), 741–1755 Hz (band 3), 1755–3758 Hz (band 4), and 3758–7741 Hz (band 5).

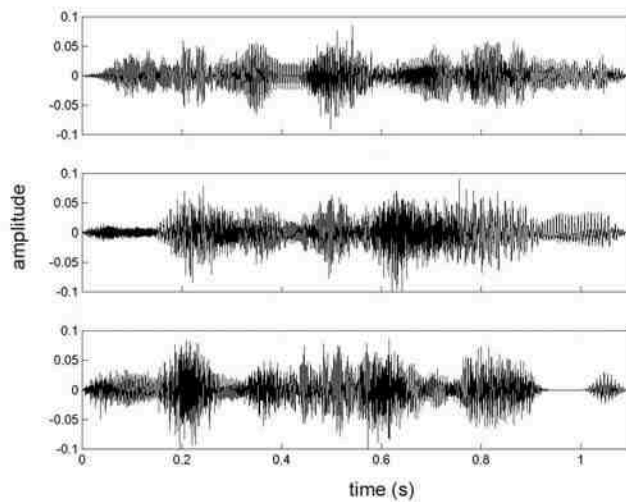


Figure 1. Example waveforms of babble signals used in this study.

bands) of 0 dB; then, based on the results of these 200 trials, the SNR was increased or decreased and the listener performed another block of 100 trials, and so forth, until he/she achieved between 60 and 70% correct responses. The resulting SNR was retained as the mid-point SNR for that listener. The noise bands were generated in the spectral domain, in such a way that the Long-Term-Average-Spectrum (LTAS) of the noise was identical to that of the speech signals. The latter LTAS was computed by averaging the power spectra of all the VCV signals used in this study. Obviously, on any given trial, the spectrum of the noise could differ substantially from the LTAS of the speech, due to the independent roving of the SNR in the different bands. The temporal waveform of the noise was obtained by combining the inverse-digital-Fourier transform of the noise spectra generated at different time points, using an overlap-add technique with zero-padding.

In the babble condition, in addition to the background noise, the target VCV signals were mixed on each presentation with a sample of multi-talker babble. Eight hundred and sixteen (816) different samples of babble were generated offline and stored on the computer hard disk. Each sample was obtained by adding 24 randomly-selected pieces of the VCV signals. (Before being added, the selected segments were given 45-ms on and off ramps, they were zero-padded on both sides to yield a duration of 1093-ms, and time-shifted to start at different delays). Example waveforms of babble samples obtained using this procedure are shown in Figure 1. Despite our use of an unusual method for generating the babble samples, the resulting signals sounded like the kinds of babble produced by several concurrent talkers. They generally contained some easily recognizable phonemes popping out in a background of virtually un-intelligible other signals that were nevertheless identified as speech by most listeners. Since the target signals presented on consecutive trials were not necessarily produced by the same speaker, no consistent speaker-related cue was available to the listeners for the

identification of the target signals. However, because the signal-to-babble ratio (SBR) was kept constant at 5 dB, the target speech signals were generally louder than the babble. Listeners were instructed to try and ignore the babble background, and repeat what they thought was the target VCV. When they heard several VCV syllables, they were asked to repeat only that which they perceived the most distinctly. In order to yield an approximately constant correct identification performance from the subjects, the mid-point SNR had to be adjusted to a lower value in the babble than in the no-babble condition (the exact percent-correct scores achieved by the listeners in these two conditions during the actual experiment are given in the Results section).

The relative weights of the different frequency bands were estimated by computing the point bi-serial correlation between the SNR for the considered band and the listener's binary identification scores (correct/incorrect) across 1500 trials [1, 2]; a response was counted as correct when the listener repeated both the consonant and the vowel correctly, and as incorrect otherwise. Doherty and Turner [1] have proposed normalizing the correlation coefficients so that their sum across bands equals one before pooling data across subjects. The theoretical justification for this normalization remains unclear. Perhaps its origin can be traced back to early work on the COSS method [14], in which the sum of the observer's weights across observations was assumed equal to unity for mathematical convenience. Later work indicating that the correlation coefficients measured using the correlational method were related to the observer's internal weights by a constant of proportionality may have been the incentive for normalizing the sum of the correlation coefficients themselves. However, it should be remarked that the assumptions of the COSS and correlational methods formalized by Berg [14], Richards and Zhu [15], and Lutfi [16] are not necessarily met by the correlational approach applied to speech by Turner and colleagues. Because of the higher spectral and temporal complexity of the speech stimuli, and the possibility of listeners employing a variety of strategies for task performance, the underlying model is likely to be more complex. Until a fully developed formal model for the multi-band speech-identification situation considered becomes available, and a clear theoretical answer to the question of whether and how to scale the obtained correlation coefficients can be produced, it is probably advisable to try different normalization procedures and to examine how this affects the conclusions. This was done in the present study. Three normalization schemes were tested; they are detailed in the following Results section, along with the corresponding results.

The subjects took part in two test sessions, which took place on different days, separated by less than seven days. Each session lasted two hours, and involved the presentation of 1500 stimuli. Eight subjects were tested first in the absence of the babble background, then in the presence of it; for the remaining 7 subjects, the order was reversed. After each stimulus presentation, the listeners had to repeat

Table I. Average nominal SNRs used and Error rates obtained in the two experimental conditions (No babble versus Babble). Four error types are considered : Overall, Consonant only, Vowel (accompanied or not with an error on the consonant) and Vowel only.

Condition	Nominal SNR (dB)	Error rate (%)			
		Overall	Consonant only	Vowel	Vowel only
No babble	-1.7 (SD=1.0)	33.6 (SD=5.1)	32.4 (SD=4.9)	5.4 (SD=1.8)	1.2 (SD=0.5)
Babble	6.3 (SD=1.5)	36.4 (SD=3.6)	34.8 (SD=3.5)	9.1 (SD=2.3)	1.6 (SD=0.6)

the target VCV. Responses were input into a computer by the experimenter.

2.3. Apparatus

Speech stimuli were acquired using a Rhode NT-1 electrostatic microphone, a Behringer ultragrain Mic 2000 preamplifier and a Turtle Beach Multisound Fiji Pro Series sound card containing a 16-bit A/D converter. The sampling frequency for acquisition and generation was 44.1 kHz. Signal processing, stimulus presentation, and response acquisition were performed using software running under Windows 98 on a Pentium III 350 MHz computer. Stimuli were presented monaurally to the subject's right ear using Sennheiser HD 265 linear II circumaural headphones, after 16-bit D/A conversion by a Roland UA30 USB audio interface. The target VCV signals were presented at 65 dB SPL peak, as measured using the slow time constant of the sound level meter. The levels of the other signals (noise and babble) were specified relative to the target signal level, as described above. Subjects were comfortably seated in a quiet room during the tests.

3. Results

3.1. Midpoint SNRs and error rates

Table I shows the average nominal SNRs used in the no-babble and babble conditions, and the error rates observed in these two conditions. The mid-point SNR was set on average 8 dB lower in the babble than in the no-babble condition. In both conditions, the observed error rates were reasonably close to the targeted 35%, indicating that the mid-point SNR was adjusted adequately. A vast majority of the errors involved consonants.

3.2. Information-theoretic analysis of the confusion matrices

Figure 2 shows, for each listener and each experimental condition, the Transmitted Information (TI) (in %) measured for five consonantal phonetic features as described in Miller and Nicely [17]. These data were analyzed using a two-way Repeated Measure Analysis of Variance (RMANOVA), with the test condition and phonetic feature as within-subject factors. Overall, the introduction of the babble produced a significant decrease in TI [main effect of condition: $F(1,14) = 33.6, p < 0.001$]. Not all features were affected [condition-by-feature interaction: $F(4,56) = 34.4, p < 0.001$]: nasality, affrication and duration suffered

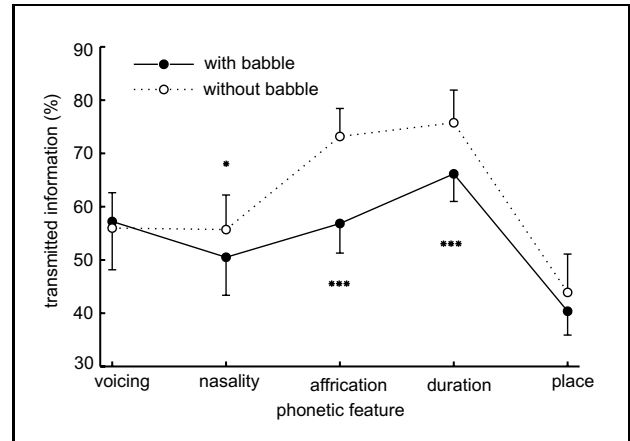


Figure 2. Amounts of transmitted information for five consonantal phonetic features in the 'babble' and 'no babble' conditions. The error bars represent standard deviations around the mean percentages of transmitted information across all listeners.

significantly (Bonferroni-corrected $p < 0.001$); voicing and place of articulation did not.

3.3. Frequency-importance functions

The psychometric functions, relating percent correct to SNR in each band, are shown in Figure 3. Although a higher-order regression model would yield better fits to the data, over the range of SNRs used here, the relationship between percent correct and SNR was in general relatively well approximated using linear fits. This approximate linearity of the psychometric functions within the considered portion of the SNR range provides some justification for the use of correlation coefficients to quantify the strength of the relationship between SNR and percent correct. Thus, we decided for simplicity and consistency reasons to follow Turner and colleagues in their use of correlation coefficients to quantify the importance of the different bands. It can be seen that the slope of the psychometric functions varied across bands, already providing a hint that SNR variations had a larger influence in some bands than in others. Finally, it can be observed too that the slopes of the psychometric functions were generally shallower in the presence than in the absence of the babble. This global effect was expected at least for two reasons: The babble was an additional source of variance in the identification scores and the babble probably limited the maximum score that can be obtained, so the range of scores was automatically limited by the presence of the babble.

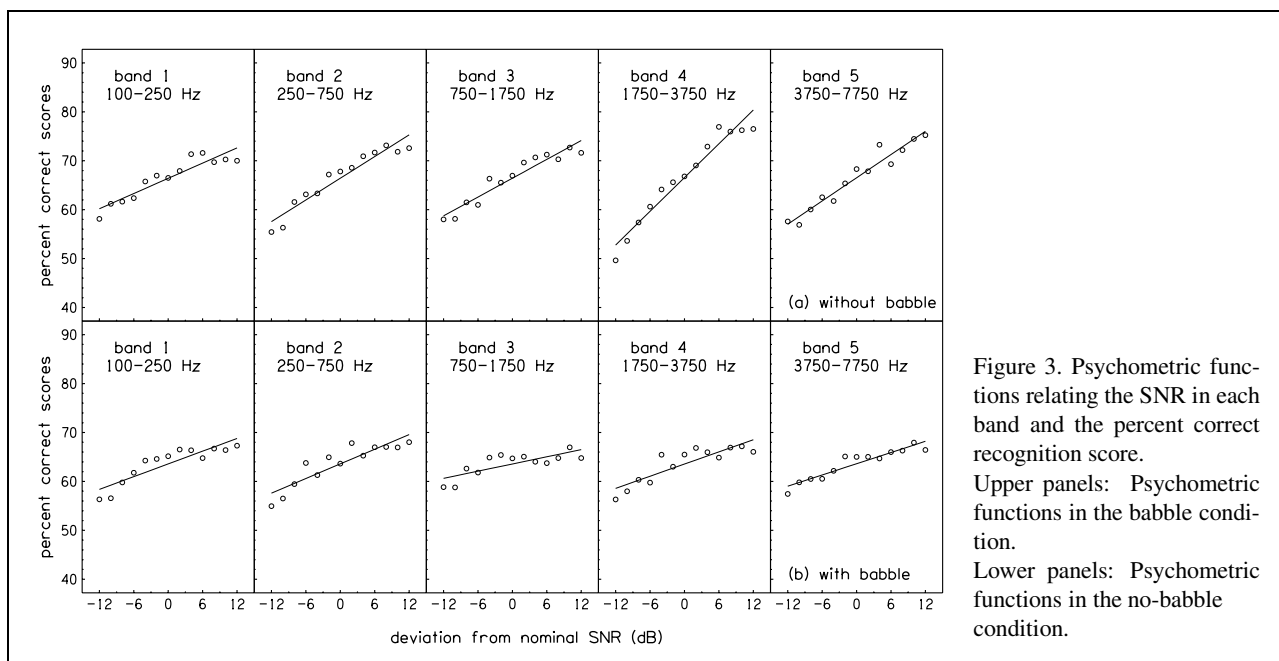


Figure 3. Psychometric functions relating the SNR in each band and the percent correct recognition score. Upper panels: Psychometric functions in the babble condition. Lower panels: Psychometric functions in the no-babble condition.

The correlation coefficients between identification scores and SNRs are given in Table II for each listener, each condition, and each frequency band. It is important to check first whether the values of these coefficients differed significantly from zero. Statistically-significant correlation coefficients ($p < 0.05$) are shown in bold. It can be seen that significant correlations were observed in a majority of cases (84% on average across all conditions, bands, and subjects). In the reference condition (i.e., without added babble), only 2 out of 75 measured correlation coefficients were not statistically different from zero. A larger number of coefficients (22 out of 75) were not significantly different from zero in the babble condition. This observation was expected because in that condition identification performance was determined not only by the SNR but also by the SBR, which was not taken into account in the calculation of the correlation coefficient. Of all the correlation coefficients measured, only one had a negative value, and this was not significantly different from zero. Thus, although negative correlation coefficients between SNR and performance could in principle be observed, due for instance to target-related information in a given band interfering with the processing of target-related information in a more important band, the present results provide no evidence for this.

Figure 4 shows the average frequency-importance functions measured in the absence and in the presence of babble. As explained above (see Methods), three different approaches for deriving average importance functions across listeners based on the individually measured correlation coefficients were tested. In the first approach, the raw correlation coefficients for each band were averaged across listeners. The average functions computed in this way are shown in the top panel of Figure 4. A comparison between the functions measured in the two test conditions reveals that the correlation coefficients measured with the

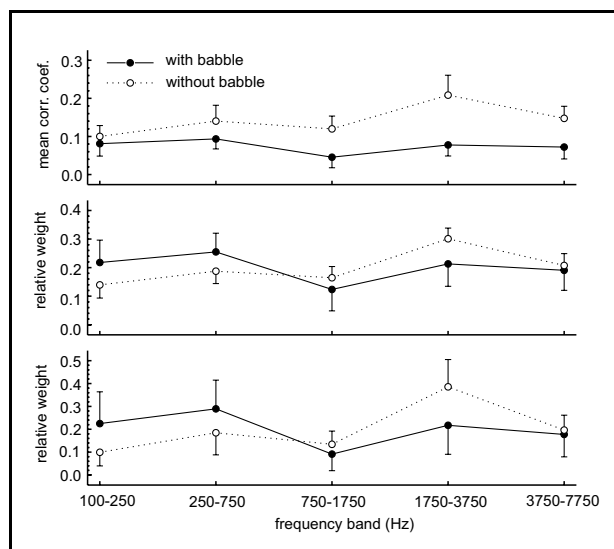


Figure 4. Frequency-importance functions measured using three different normalization schemes in the absence and in the presence of the babble background. The error bars represent standard deviations around the mean relative weights for the considered frequency band across all listeners. Upper panel: Raw correlation averaged across listeners (No normalization at all). Middle panel: Normalization applied to the correlation coefficients in each band so that their sum equals 1. Lower panel: Normalization applied to the squared correlation coefficients (R^2) in each band so that their sum equals 1.

babble present were usually reduced compared to those measured without the babble, especially on the higher frequency bands where the average coefficients were larger. The most dramatic reduction was observed for band 4 (1750-3750 Hz), for which the average correlation coefficient measured in the reference (no-babble) condition was the largest. These observations were confirmed by the re-

sults of a RMANOVA with frequency bands and listening condition as within subject factor, and the testing order as between subject factor. This RMANOVA showed a significant overall difference between the correlation coefficients measured in the two conditions [$F(1,13) = 202.1$, $p < 0.001$], the difference depending upon the frequency band [$F(4,52) = 12.6$, $p < 0.001$].

The second approach that was used to derive the average importance functions was inspired by Doherty and Turner [1] and Turner *et al.* [2]. These authors scaled the correlation coefficients so that their sum across bands was always equal to one within each listener before averaging across listeners. As pointed out above, the theoretical justification for this normalization is unclear. However, it can be thought of, simply, as an attempt to capture the relative importance of different frequency bands, as opposed to their absolute importance. From that point of view, it can be seen that when the babble was introduced, the lowest two frequency bands acquired more weight whilst higher frequency bands, especially the fourth, became relatively less important. These observations were confirmed by the results of a RMANOVA using the same factors as precedingly. This RMANOVA showed a significant across frequency bands difference [$F(4,52) = 10.7$, $p < 0.001$], as well as a significant interaction between the frequency band factor and the condition (babble vs no babble) [$F(4,52) = 9.5$, $p < 0.001$]. Post-hoc pairwise comparisons were performed in order to determine between which bands the normalized correlation coefficients differed significantly, for each condition separately. The results showed that in the absence of the babble, the normalized correlation coefficients for band 4 (1750-3750 Hz) were on average significantly larger than for the other bands ($6.257 < t < 9.4$, $df = 14$, with Bonferroni-corrected p values always < 0.001). In the presence of the babble, the only significant difference was between bands 2 (250-750 Hz) and 3 (750-1750 Hz) ($t = 4.4$, $df = 14$, Bonferroni corrected $p < 0.01$). Pairwise comparisons also confirmed the above observation of a shift in the importance functions in favor of low frequencies when the babble was introduced: significant differences in the normalized correlation coefficients were observed between the babble and no-babble conditions for bands 1 ($t = -3.5$, $df = 14$, $p < 0.05$), 2 ($t = -3.3$, $df = 14$, $p < 0.05$), and 4 ($t = 4.8$, $df = 14$, $p < 0.01$).

The same general trends were observed when, instead of normalizing the correlation coefficients so that their sum across bands was constant, the squares of the correlation coefficients were normalized so that the sum across bands was constant in all listeners and both conditions (lower panel in Figure 4). The squared correlation coefficient, R^2 , can be thought of as the proportion of the variance in a dependent variable that is accounted for by an independent variable in a linear regression model. If the model is hypothesized to be additive, the total variance in the response scores that is explained by SNR variations is simply equal to the sum of the variances in response scores that are explained by SNR variations within the different

Table II. The point biserial correlations for the five bands of speech for the $N=15$ individual normal hearing listeners. The signal-to-noise ratio in each band was correlated with the listener's responses (correct=0 versus incorrect=1) from the trial by trial experimental record. The bold entries indicate correlation coefficients whose absolute value exceeded the upper limit of the 95% confidence interval around zero, i.e., $|r| > 1.96/\sqrt{1500}$ ($=0.0506$), where 1.96 is the z (inverse cumulative normal) score corresponding to the 95% confidence interval, 1500 is the number of observations on which the estimate of the true correlation coefficient is based. This formula assumes that the sampling distribution of the estimate is normal, an assumption which can be considered reasonable for large N [16].

Without babble					
Subject	band 1	band 2	band 3	band 4	band 5
GG	0.08	0.09	0.05	0.23	0.14
FA	0.08	0.08	0.10	0.19	0.12
GM	0.10	0.05	0.04	0.17	0.14
LA	0.08	0.15	0.12	0.22	0.16
SO	0.11	0.12	0.10	0.15	0.13
YA	0.04	0.11	0.12	0.19	0.17
AU	0.08	0.08	0.10	0.17	0.13
JO	0.05	0.13	0.11	0.17	0.11
AX	0.09	0.11	0.08	0.15	0.11
BE	0.11	0.10	0.13	0.17	0.09
CA	0.13	0.13	0.12	0.15	0.13
CG	0.11	0.11	0.09	0.15	0.06
CR	0.07	0.21	0.10	0.20	0.15
SA	0.05	0.14	0.13	0.21	0.11
DA	0.07	0.11	0.14	0.22	0.16
With babble					
Subject	band 1	band 2	band 3	band 4	band 5
GG	0.09	0.07	0.01	0.09	0.09
FA	0.04	0.09	0.04	0.05	0.06
GM	0.07	0.07	0.04	0.09	0.01
LA	0.06	0.12	-0.01	0.07	0.09
SO	0.03	0.08	0.06	0.07	0.07
YA	0.05	0.05	0.06	0.09	0.07
AU	0.02	0.07	0.06	0.06	0.06
JO	0.05	0.05	0.01	0.05	0.02
AX	0.12	0.05	0.06	0.05	0.06
BE	0.08	0.11	0.06	0.02	0.06
CA	0.08	0.07	0.06	0.03	0.04
CG	0.07	0.10	0.03	0.06	0.08
CR	0.11	0.09	0.02	0.09	0.03
SA	0.07	0.06	0.04	0.10	0.08
DA	0.08	0.08	0.05	0.04	0.09

bands. In this context, keeping the sum of the R^2 s across bands constant across listeners and conditions appears like an appropriate way of compensating for global differences in the proportion of variance in response scores (explained by SNR variations) between listeners and conditions. Perhaps the mathematical basis of this normalization is easier to grasp than that of the normalization of the correlation coefficients used by Turner and colleagues. However, the important point is that whichever of the two normalization procedures is used, the main trends in the results

remain unchanged: in the presence of the babble, lower frequencies (below about 750 Hz) become relatively more important while higher frequencies become relatively less important.

4. Discussion

4.1. Result summary

The main result of the present study is that the shape of frequency-band importance functions for the identification of nonsense syllables, as measured using the correlational approach of Turner and colleagues, was significantly affected by the presence of a competing multi-talker babble background. Whereas in the absence of the competing babble, a frequency band corresponding to middle-to-high frequencies (1750-3750 Hz) made the most important contribution to correct identification of the speech material, in the presence of the babble, this band lost most of its importance. Other frequency bands (the lowest two bands which spanned 100-750 Hz) were less affected, so that, relatively to other bands, they became more important in the presence of the babble than they were in its absence. This relative increase was visible only after the correlation coefficients (or R^2) between SNR and identification scores were normalized in such a way that their sum across bands no longer differed between the two conditions. The increase was not apparent when looking at the raw correlation coefficients, which generally decreased after the babble was introduced.

4.2. Limitations of the present study

It is important to acknowledge some limitations of the present study. First, frequency importance functions estimated using the correlational method may depend upon how the spectrum is partitioned. In [2], the frequency bands were selected in such a way that their importance, as measured using AI importance functions, was roughly equal. Accordingly, under the hypothesis that there would be no interactions between the bands when they were presented together, flat frequency importance functions were expected and significant deviations from this pattern could perhaps be interpreted as reflecting the listeners' frequency-weighting strategies. In the present study, the different bands were not equated for AI importance, and different weights across bands do not necessarily reflect differences in the listener's weighting of information across bands. Here, different weights across bands may simply reflect the fact that some of the bands intrinsically contain more useful information for speech identification than others. For instance, the estimated weights may differ across bands even if listeners are paying equal attention to all bands. Thus, it is important not to confuse the "weights" estimated here with observer's weights defined in earlier articles on the correlational method, such as [14] or [15]. Similarly, the differences in relative weights between the two test conditions (with and without babble) cannot be interpreted unequivocally in terms of changes in the listener's strategies: the fact that some bands became

relatively more important, and others less important, when the babble background was introduced may reflect the way in which the spectral distribution of cues for speech identification was altered by the babble. Note that this limitation would not have been alleviated by selecting the bands to yield equal AI importance because the introduction of the babble is likely to alter the relative importance of the different bands in a way that is not well accounted for by the AI. More fundamentally, one of the main ideas behind the application of the correlational approach to speech perception is that the AI model does not account for interactions between simultaneously-presented bands; thus, in the context of that approach, selecting bands based on AI considerations may be construed as self-contradictory.

A second point that must be kept in mind is that the frequency-importance functions measured here, like those measured by Doherty and Turner [1] and Turner *et al.* [2], or the importance functions of the AI or SII, reflect the general or "average" importance of different frequency bands for speech identification. The finding of a relatively large weight for a given frequency band does not imply that this band is important for identifying all of the presented speech signals, nor that it is dominant at all times whilst the signal is presented. Speech signals are highly dynamic, and important cues for their identification are likely to occupy different frequency regions at different times. Furthermore, listener's attention to each frequency band may vary rapidly over time to accommodate such spectral changes. The notion of average differences in importance across frequency bands is supported by numerous earlier results in the speech perception literature, including in particular those that led to the formulation of the AI and SII.

A third point worth acknowledging is that, with the approach used here, the influence of interfering babble on the identification of target speech signals was measured indirectly, via changes in the effect of an external noise on target speech identification. The effect of the babble in a given band was probably influenced by the level of the noise in that band. In practice, the fact that the babble was generated from the target speech and that it had the same LTAS and an RMS level only slightly lower reduces the likelihood that the babble was affected very differently by the noise than the target speech, across the different bands.

4.3. Possible reasons for the different effects of the babble at high and low frequencies

A first possible reason why the measured importance of lower-frequency bands was proportionally less reduced when the babble was introduced than that of higher frequency bands is simply that the SBR was perhaps lower in the latter than in the former bands. This could have been the case if for some reason the babble contained more energy at high frequencies than at low frequencies compared to the signal. In order to check this possibility, we plotted the LTAS of the babble and target speech signals as well as the corresponding SBR across frequency bands (Figure 5; the SBR values for each band are shown under the top

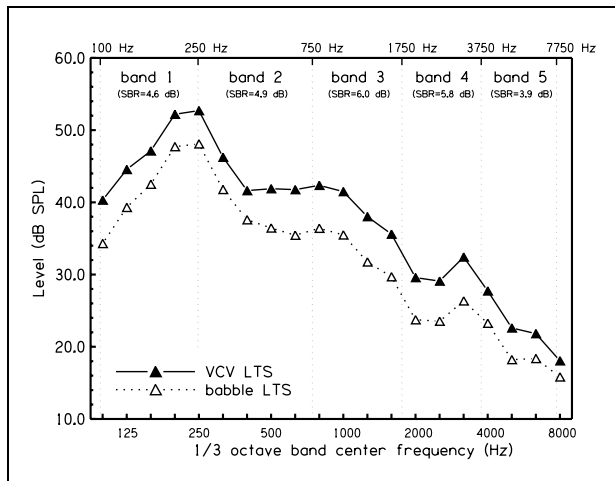


Figure 5. Long term average spectra of the target speech and babble stimuli used in this study.

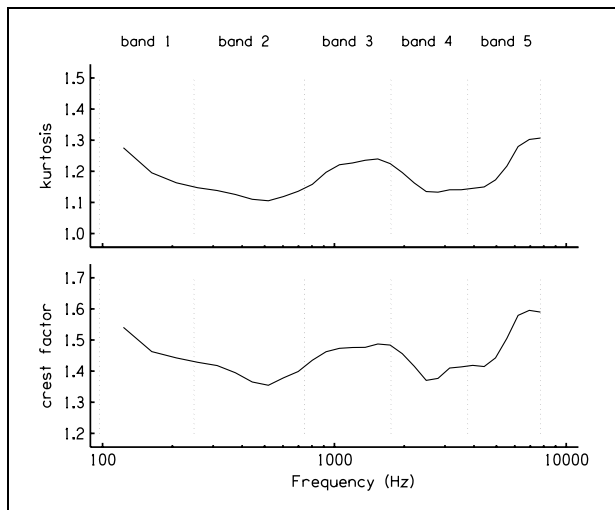


Figure 6. Quantification of the amount of envelope temporal fluctuations with two metrics. Upper Panel: The fourth moment (or kurtosis) of the envelope divided by the square of the envelope power [18]. Lower panel: The envelope maximum divided by the rms power called the crest factor [18]. The envelopes were extracted by full wave rectifying and then low-pass filtering (filter type: 2nd order butterworth, cutoff frequency = 50 Hz) the output of Gammachirp filters [19].

axis). The LTAS of the babble was very similar in shape to that of the target signals, and the SBR was, if anything, slightly larger – i.e., more favorable – in the third and fourth bands than in the first and second bands. Thus, the relative increase in the importance of the latter two bands in the presence of the babble cannot be explained simply by differences in the long-term spectral characteristics of the target and babble signals.

Speech is a dynamic signal. Thus, even though the SBR was approximately constant across bands in the long run, it could nevertheless fluctuate from moment to moment. It has been suggested that when trying to identify speech signals in the presence of maskers whose temporal envelope fluctuates, normal-hearing listeners take advantage of dips

in the masker envelope in order to extract the target signal – a form of selective listening in the temporal domain often referred to as “listening in the valleys” (e.g. [20, 21, 22]). If, for some reason, the babble signals used in this study contained more marked envelope fluctuations at low than at high frequencies, this might explain the relative increase in the importance of low frequencies in the presence of the babble. In order to test this hypothesis, we measured the average amount of temporal envelope fluctuations in the simulated responses of auditory filters (Gammachirp; [19]) in response to the babble signals. Two standard indices were used [18]: the kurtosis or fourth moment of the envelope divided by the square of the envelope power, and the envelope crest factor, defined as the envelope maximum divided by the RMS. As can be seen in Figure 6, neither of the indices used provide evidence that the relative size of the fluctuations in the envelope of peripheral auditory-filter responses to the babble were larger at lower than at higher frequencies. This observation is inconsistent with an explanation in terms of “dip listening” for our finding that in the presence of the babble, lower frequencies became relatively more important than high ones.

Other possible explanations for this finding are variations on the idea that the auditory system is somehow better at extracting the information in concurrent signals at low than at high frequencies. Two functional characteristics of the peripheral human auditory system that are thought to play a crucial role in the segregation of simultaneous signals are frequency resolution and phase locking. Both are known to decrease with increasing frequency. Accurate phase locking may be an important factor in the ability to segregate simultaneous periodic or quasi-periodic signals having different periodicities, such as concurrent vowels uttered by different speakers. Indeed, although various schemes have been proposed for how the components of such simultaneous periodic signals could be analyzed based either on the fine structure or on the envelope temporal information encoded in auditory-nerve responses (e.g. [23]), the periodicity information conveyed by envelope fluctuations may be weak when the phases of the frequency components are not such that they ensure marked envelope fluctuations in the auditory filter outputs. In these circumstances, accurate phase locking and temporal fine structure information may be required in order for temporal-based separation schemes to work.

As regards frequency resolution, although some results in the literature indicate that fine frequency resolution is not necessarily required for accurate speech recognition (e.g. [24]), these results apply primarily to quiet listening situations. In the presence of background noise or competing speech, access to detailed spectral information appears to be important for speech recognition [25, 26, 27, 28], especially when the masker is fluctuating [29]. A possible reason for this increased importance of frequency resolution for speech identification in masked situations is that, for a speech signal to be correctly identified, it may be necessary first to segregate the frequency components that correspond to the signal from those that pertain to

the masker. Unless there is sufficient temporal information in peripheral auditory-nerve responses for analyzing the frequency components present, the ability to segregate these components will depend on cochlear frequency resolution. Although frequency resolution, when measured as auditory-filter bandwidth divided by center frequency, actually increases slightly with increasing center frequency, the auditory-filter bandwidth itself increases substantially toward high frequencies. Thus, if the average frequency spacing in Hz between the signal and masker components is not systematically larger at high than at low frequencies (which is likely to be the case for speech and most other natural signals, especially when these are harmonic), the signal and masker components are more likely to interact in the auditory-filter passbands at high than at low frequencies. Thus, if the frequency resolvability of components in concurrent signals effectively plays a role in the perceptual separation of these signals and their later identification, the reduced frequency resolvability of components toward higher frequencies may explain the present finding that, in the presence of babble, low frequencies became relatively more important.

4.4. Relationships between the changes in the importance functions and in the phonemic confusions patterns upon the introduction of the babble

Besides altering the relative importance of the different frequency bands, the introduction of the competing babble induced some changes in the nature of the identification errors. In particular, the percentage of TI for the affrication feature was significantly reduced by the introduction of the babble. Affrication is known to be associated acoustically with the presence of energy at relatively high-frequencies. For example, for male speakers, sibilant consonants are characterized acoustically by sharp high-frequency peaks centered at 2.5-3 kHz for the *ʃ* and *ʒ* palatals, or 4 kHz for the *ʃ* and *z* alveolars [30, 31, 32]. Therefore, a lower weighting of information at frequencies between roughly 1.8 and 3.8 kHz (band 4), as observed in the babble condition, may understandably be associated with a lower identification performance for affricative consonants. In contrast, the introduction of the babble had little influence on the identification of voicing. This resilience of voicing information to competing speech may be tentatively related to the observation that in the presence of the babble, the relative weight of low-frequency bands (<750 Hz) was increased. At this stage however, this and other possible relationships between the observed changes in the shape of the importance functions and the changes in the patterns of phonemic confusions remain highly speculative. This appears like something worth trying to elucidate in future studies.

5. Conclusions

Frequency-importance functions for the identification of VCV syllables by normal-hearing listeners were measured successively in the absence and in the presence of multi-talker babble, using a correlational approach. The shapes

of the importance functions measured in the two conditions were found to be significantly different: whilst the importance of the middle-to-high frequency band (1750-3750 Hz) was significantly reduced by the babble, relatively to this band and the others, the importance of the lowest two frequency bands (100-750 Hz) was increased in the presence of the babble, compared to when the babble was absent. An analysis of the long-term average spectral characteristics of the babble and of the fluctuations in its temporal envelope after peripheral filtering provided no obvious reason for the relative increase in the importance of the lowest two frequency bands and concomitant decrease in the importance of the middle-to-high frequency band. A possible explanation for this finding, which deserves further scrutiny, is that the auditory system is better at extracting speech from concurrent speech in lower frequency bands than in higher ones, possibly due to finer encoding of spectral and temporal fine structure information at lower frequencies.

Acknowledgements

The authors are grateful to Fabien Masquelier and Guillaume Morel for their help in running the experiments. This research was supported by a research grant from the Ministère de l'Éducation Nationale et de la Recherche awarded to the first author, and was performed in the framework of the Groupe de Recherche GDR 2213 "Prothèses auditives" linking the CNRS (UMR CNRS 5020, Pr. Collet) to CCA Groupe, Entendre, Oticon, Phonak, and Siemens Audiologie. The authors wish to thank Pierre Badin, Brian Moore, Andrew Oxenham and an anonymous reviewer for their helpful comments.

References

- [1] K. A. Doherty, C. W. Turner: Use of a correlational method to estimate a listener's weighting function for speech. *J. Acoust. Soc. Am.* **100** (1996) 3769–3773.
- [2] C. W. Turner, B. J. Kwon, C. Tanaka, J. Knapp, J. L. Hubbart, K. A. Doherty: Frequency weighting functions for broadband speech as estimated by a correlational method. *J. Acoust. Soc. Am.* **104** (1998) 1580–1585.
- [3] N. R. French, J. C. Steinberg: Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* **19** (1947) 90–119.
- [4] H. Fletcher: *Speech and hearing in communication*. Krieger, New York, 1953.
- [5] K. D. Kryter: Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.* **34** (1962) 1689–1697.
- [6] C. V. Pavlovic, G. A. Studebaker, R. L. Sherbecoe: An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. *J. Acoust. Soc. Am.* **80** (1986) 50–57.
- [7] A. S3.5-1997: American national standards methods for the calculation of the speech intelligibility index. ANSI, New York, 1997.
- [8] R. M. Warren, K. R. Riener, J. Bashford, J. A., B. S. Brubaker: Spectral redundancy: intelligibility of sentences

- heard through narrow spectral slits. *Percept. Psychophys.* **57** (1995) 175–182.
- [9] R. P. Lippman: Intelligibility of bandpass speech: Effects of truncation or removal of transition bands. *IEEE Trans. Speech Audio Process.* **4** (1996) 66–69.
- [10] K. M. Grant, L. D. Braida: Evaluating the articulation index for auditory-visual input. *J. Acoust. Soc. Am.* **89** (1991) 2952–2960.
- [11] H. Müsch, S. Buus: Using statistical decision theory to predict speech intelligibility. I. Model structure. *J. Acoust. Soc. Am.* **109** (2001) 2896–2909.
- [12] A. W. Bronckorst, R. Plomp: Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *J. Acoust. Soc. Am.* **92** (1992) 3132–3139.
- [13] D. S. Brungart: Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* **109** (2001) 1101–1109.
- [14] B. G. Berg: Analysis of weights in multiple observation tasks. *J. Acoust. Soc. Am.* **86** (1989) 1743–1746.
- [15] V. M. Richards, R. Zhu: Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *J. Acoust. Soc. Am.* **95** (1994) 423–434.
- [16] R. A. Lutfi: Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks. *J. Acoust. Soc. Am.* **97** (1995) 1333–1334.
- [17] G. A. Miller, P. E. Nicely: An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* **27** (1955) 338–352.
- [18] W. M. Hartmann, J. Pumplin: Noise power fluctuations and the masking of sine signals. *J. Acoust. Soc. Am.* **83** (1988) 2277–2289.
- [19] T. Irino, R. D. Patterson: A time-domain level dependent auditory filter: The gammachirp. *J. Acoust. Soc. Am.* **101** (1997) 412–419.
- [20] P. A. Howard-Jones, S. Rosen: The perception of speech in fluctuating noise. *Acustica* **78** (1993) 252–272.
- [21] S. Arlinger, H. A. Gustafsson: Masking of speech by amplitude modulated noise. *Journal of Sound and Vibration* **151** (1991) 441–445.
- [22] H. A. Gustafsson, S. D. Arlinger: Masking of speech by amplitude modulated noise. *J. Acoust. Soc. Am.* **95** (1994) 518–529.
- [23] P. A. Cariani: Neural timing nets. *Neural Netw.* **14** (2001) 737–753.
- [24] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, M. Ekelid: Speech recognition with primarily temporal cues. *Science* **270** (1995) 303–304.
- [25] T. Baer, B. C. Moore: Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *J. Acoust. Soc. Am.* **95** (1994) 2277–2280.
- [26] Q. J. Fu, S. R. V., W. X.: Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. *J. Acoust. Soc. Am.* **104** (1998) 3586–3596.
- [27] M. F. Dorman, P. C. Loizou, J. Fitzke, Z. Tu: The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear implant signal processors with 6–20 channels. *J. Acoust. Soc. Am.* **104** (1998) 3583–3585.
- [28] L. M. Friesen, R. V. Shannon, D. Bazkent, X. Wang: Speech recognition in noise as a function of spectral channels: Comparison of acoustic hearing and cochlear implants. *J. Acoust. Soc. Am.* **110** (2001) 1150–1163.
- [29] M. K. Qin, A. J. Oxenham: Effect of simulated cochlear implant processing on speech reception in fluctuating maskers. *J. Acoust. Soc. Am.* **114** (2003) 446–454.
- [30] P. Stevens: Spectra of fricative noise in human speech. *Lang. Speech* **3** (1960) 32–49.
- [31] W. Jassem: Formants of fricatives consonants. *Lang. Speech* **8** (1965) 1–16.
- [32] S. J. Behrens, S. E. Blumstein: Acoustic characteristics of English voiceless fricatives: A description analysis. *J. Phonetics* **16** (1988) 295–298.