Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation

Mickael L. D. Deroche[a)]
Cochlear Implants and Psychophysics Lab, Department of Hearing and Speech Sciences, University of Maryland, College Park, Maryland 20742

John F. Culling
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom

(Received 16 May 2011; revised 26 August 2011; accepted 31 August 2011)

Two experiments investigated listeners' ability to use a difference of two semitones in fundamental frequency (F0) to segregate a target voice from harmonic complex tones, with speech-like spectral profiles. Masker partials were in random phase (experiment 1) or in sine phase (experiment 2) and stimuli were presented over headphones. Target's and masker's harmonicity were each distorted by F0 modulation and reverberation. The F0 of each source was manipulated (monotonized or modulated by 2 semitones at 5 Hz) factorially. In addition, all sources were presented from the same location in a virtual room with controlled reverberation, assigned factorially to each source. In both experiments, speech reception thresholds increased by about 2 dB when the F0 of the masker was modulated and increased by about 6 dB when, in addition to F0 modulation, the masker was reverberant. Masker partial phases did not influence the results. The results suggest that F0-segregation relies upon the masker's harmonicity, which is disrupted by rapid modulation. This effect is compounded by reverberation. In addition, F0-segregation was found to be independent of the depth of masker envelope modulations. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3643812]

PACS number(s): 43.66.Dc, 43.55.Hy [CJP] Pages: 2855–2865

I. INTRODUCTION

It is known that a difference in fundamental frequency (ΔF0) between simultaneous speech messages facilitates intelligibility of a target voice. Brokx and Nooteboom (1982) resynthesized speech recordings of two voices from a linear predictive coding analysis, so that they controlled the fundamental frequency (F0) contour of each sentence. Whether those voices were monotonized or intonated, words spoken by competing voices with different F0s (or different mean F0s for the intonated voices) were reported more accurately than those spoken with the same F0. The larger the ΔF0, the lower the percentage in errors in reporting words, except in the monotonized case when the ΔF0 equalled one octave. These results led to the idea that harmonicity must be involved in the segregation of concurrent voices by F0.

The "voiced" parts of speech, which are harmonic, are the vowels and the sonorant consonants such as /m/, /w/, and /r/. Simultaneous competing vowels were consequently often chosen as experimental stimuli to investigate the role of harmonicity underlying the ΔF0 effect. Nevertheless, the improvement in recognition with ΔF0 is somewhat different for steady-state vowels than for spoken speech. In several experiments, recognition of simultaneous vowels increased when ΔF0 increased up to about one semitone and asymptoted for larger ΔF0s (Scheffers, 1983; Summerfield and Assmann, 1991; Culling and Darwin, 1993).

A. Models of ΔF0 effect

The mechanisms underlying the ΔF0 effect have been a matter of controversy. One approach was a strategy guided by the identification of competing F0s (Scheffers, 1983; Assmann and Summerfield, 1990). Whether F0s were identified via a place mechanism, like a harmonic sieve (Parsons, 1976; Scheffers, 1983; Assmann and Summerfield, 1990), or via a place-time mechanism, like autocorrelation (Licklider, 1951; Assmann and Summerfield, 1990), competing F0s were both identified and vowels were then classified by a template-matching procedure. The performance of place models depended critically on the resolution of spectral analysis. Frequency selectivity of the peripheral auditory system estimated by Moore and Glasberg (1983) did not appear to be sufficiently fine for such models to predict accurately the data on ΔF0 effects. Consistently, Assmann and Summerfield (1990) showed that place-time models were better than place models at predicting the data, but still failed to show progressive improvement in identification with ΔF0.

In a second approach, also using autocorrelation, channels were segregated into two groups on the basis of the F0 of the first vowel only; the second vowel was identified from all remaining channels (Meddis and Hewitt, 1992). It might take some time for the auditory system to perform this channel separation, which might explain a smaller improvement in identification for 50-ms than for 200-ms double-vowel stimuli (Assmann and Summerfield, 1990, 1994, Culling and Darwin, 1993). This channel separation procedure succeeded in predicting the progressive improvement with ΔF0.

The idea that listeners could switch from one subset of harmonics to another by somehow inhibiting a dominant set

a)Author to whom correspondence should be addressed. Electronic mail: mderoche@hesp.umd.edu

led to a third approach. In the classic double-vowel paradigm, the two vowels were mutually masking each other and listeners were asked to report the two vowels correctly. The question arose as to whether harmonicity of the target vowel (harmonic enhancement), or the interfering vowel (harmonic cancellation) or both, underpinned the ΔF0 effect. Two experiments (de Cheveigné *et al.*, 1995; Summerfield and Culling, 1992) showed that it made no difference whether a target was harmonic or inharmonic, but performance was much better if the interfering vowel was harmonic than if it was inharmonic. In a similar approach, Lea (1992) showed that a noise-excited vowel was more accurately identified than a harmonic vowel when they were presented simultaneously. The auditory system thus appears to segregate vowels by exploiting the harmonic structure of the interfering vowel in order to suppress it, the remaining vowel becoming more intelligible through the removal of this interfering vowel. This idea has been formalised as the harmonic cancellation mechanism (de Cheveigné *et al.*, 1997). The improvement with ΔF0 of the identification of weak targets, at target-to-masker ratio (TMR) up to −20 dB, was consistent with such a process, since the estimation of the target's F0 is made difficult while that of the interferer is facilitated.

The present experiments investigated further whether harmonicity of the target source or harmonicity of the competing source was most relevant in F0-guided segregation occurring in cocktail-party situations (Cherry, 1953). Since an inharmonic voice is a highly artificial stimulus, inharmonicity was produced by simulating natural environments where voices are no longer strictly harmonic: F0 modulation can blur harmonicity and reverberation exacerbates this effect, as explained below.

B. Detrimental effect of F0 modulation

The cancellation mechanism must have a limited temporal resolution beyond which dynamic harmonic masking stimuli cease to be effectively cancelled. In the spectral domain, the effect of F0 modulation would be to blur the representation of the F0 in the masker's excitation pattern. In the time domain, there must be some finite time window over which neural discharges are integrated, and F0 movement during this window will distort the information therein. In either case, the auditory system may require some time to identify the masker's F0 such that the tuning of the cancellation mechanism lags the actual F0 at any given moment. When the masker's periodicity changes over time, cancellation should thus be suboptimal to some extent. Note that cancellation may also be useless if the stimulus has stopped by the time the masker periodicity has been identified, thereby also accounting partly for the duration effect observed in double-vowels (McKeown and Patterson, 1995). Thus, when the rate of F0 modulation exceeds the temporal resolution of cancellation, the stimuli cannot be cancelled as effectively as they would be if they were steady. Culling *et al.* (1994) measured double-vowels recognition and showed that F0 modulation of ±2 semitones at 5 Hz reduced the ΔF0 benefit by 6 dB in anechoic conditions (experiment 3).

C. Detrimental effect of reverberation combined with F0 modulation

Reverberation adds delayed copies of the direct sound. The reflections are delayed by their path between walls of the room, so reflected sounds always arrive later than the direct sound. If the F0 is constant over time, the reflections bring the same F0 as the direct sound, but if the F0 varies over time, then the listener's ear simultaneously receives the F0s of the direct sound and of the various reflections. Harmonic cancellation would presumably suffer from the presence of a masker with several F0s. In Culling *et al.* (1994), the ΔF0 benefit was reduced by 10 dB in reverberant compared to anechoic conditions for an F0 modulation of ±2 semitones at 5 Hz, while reverberation had no effect when vowels were monotonized (experiment 3).

D. The present experiments

In Culling *et al.* (1994), F0 modulation of the two vowels was varied together, as was reverberation, leaving it uncertain whether this effect was due to modulation of the target's F0, of the masker's F0 or of both. Thus, the first aim of the present study was to determine whether segregation of a voice by F0 relied primarily on harmonicity of the target voice or on harmonicity of the masker. Given the results aforementioned (Lea, 1992; de Cheveigné *et al.*, 1995, 1997), harmonicity of the masker was expected to be most relevant. The second experiment replicated the design of the first, but changed the phase relationships between masker partials from random to sine phase. The phase relationships between partials of a complex can dramatically change the outputs of broad basal filters in which many partials interact, but changes little the output of apical filters resolving individual partials. Therefore, discrepancies in the results of the two experiments would be informative regarding the relative roles of spectral regions in the expected effect of masker's harmonicity.

II. GENERAL EXPERIMENTAL METHOD

A. Listeners

Sixteen listeners took part in experiment 1 and 16 different listeners took part in experiment 2. They were all undergraduate students, aged between 20 and 30 years old, who were paid for their participation. All listeners reported normal hearing and English as their first language. None of them were familiar with the sentences used during the test. Each listener attended a single experimental session that lasted between 60 and 80 minutes, depending on how fast the listener typed his responses.

B. Stimuli and conditions

Depending on the type of stimuli used in the literature, several other mechanisms have been shown to contribute to the ΔF0 effect. The improvement in vowel recognition occurs for such small values of ΔF0 that waveform interactions due to the beating of close partials can play an additional role (Assmann and Summerfield, 1990, 1994; Culling

and Darwin, 1994). The choice of speech as target stimuli was expected to remove any role played by waveform interactions in the present experiments. In addition, when a target sentence is masked by another sentence, listeners might confuse which sentence they should listen to, and switch between them, a form of informational masking. However, listeners are very good at using a variety of cues to overcome this attentional problem, an ability termed streaming. The choice of complex tones as maskers was expected to prevent streaming by F0 under the assumption that speech and tone were sufficiently different to be unconfusable.

The maskers were created from broadband random-phase (experiment 1) or sine-phase (experiment 2) harmonic complexes, based on a 110-Hz F0. The F0 was either fixed or sinusoidally modulated by ±2 semitones at 5 Hz. For the fixed F0 condition, the monotonized complex was filtered with a linear-phase FIR filter designed to match the average excitation pattern of 16 sentences, monotonized at 110 Hz (the masker F0). For the modulated F0 condition, the complex was filtered with a FIR filter designed to match the average excitation pattern of the 16 sentences, F0-modulated around 110 Hz. The spectral profile of the monotonized complex was similar to the excitation pattern of a single monotonized sentence, except that it was shifted two semitones lower. The spectral profile of the F0-modulated complex had smoother peaks due to modulation of the harmonic structure averaged over time. The presence of low-order partials resulted in a salient pitch. For convenience, this speech-shaped harmonic complex is hereafter referred to as "buzz."

The corpus of sentences comes from the Harvard Sentence List (IEEE, 1969). The anechoic recordings of the male voice DA, made at MIT, were used as the basis of all target stimuli. The sentences have low predictability and each has five keywords which we highlight with capitals. For instance, one sentence used in the current experiment was "the PEARL was WORN in a THIN SILVER RING." The sentences were manipulated using the PRAAT PSOLA speech analysis and resynthesis package, which calculated the F0 contour for each sentence and resynthesized the sentence with a specified F0 throughout. The mean F0 of the target sentences was higher than that of the maskers by two semitones (123.5 Hz). The modulation widths of the target sentences were 0 or ±2 semitones. F0 modulation was in phase with that of the buzz maskers when they occurred together (in 4 of 16 conditions). All maskers were longer than all target sentences so that every target word was potentially masked. Onset asynchrony (Darwin and Ciocca, 1992; Ciocca and Darwin, 1993) is known to be a powerful cue to auditory grouping and so will contribute to the perceptual segregation of speech from buzz. The onsets of the masking complexes preceded those of the speech only by the leading silence left after editing of the speech stimuli; the differences were mainly in offset. Nonetheless, this cue will occur similarly for all experimental conditions, leaving the effects observed between conditions unaffected. The monotonized speech sounded like a robotic voice, whereas the F0-modulated speech sounded rather like an old man's voice. Thus, both F0 manipulations disrupted the normal intonation contour of the original sentences.

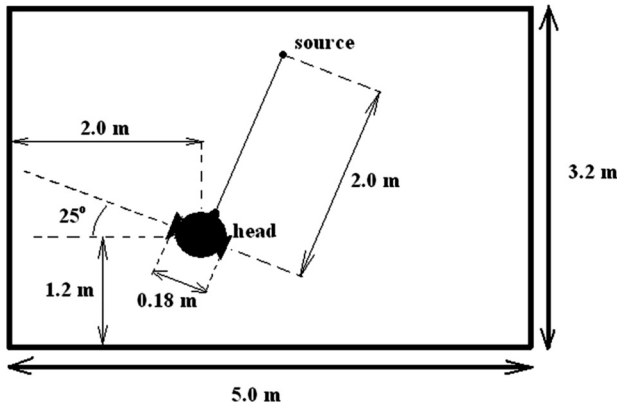

FIG. 1. Spatial configuration and virtual room considered in the two experiments.

Reverberation was added using the image (ray-tracing) method (Allen and Berkley, 1979; Peterson, 1986) as implemented in the IWAVE signal processing package (Culling, 1996). The virtual room and source/receiver configuration was identical to that of Culling *et al.* (1994). The room had dimensions 5 m long × 3.2 m wide × 2.5 m high and virtual sources were 2 m from the receivers (Fig. 1). The two receivers, separated by 18 cm, were placed along an axis rotated at 25° from the plane parallel to the 5-m wall, on either side of a center point located 1.2 m from the 5-m wall and 2 m from the 3.2-m wall. Reverberation adds irregular perturbations to the stimulus spectrum, known as room coloration. These perturbations were removed using a further FIR filter as part of a package of energetic equalization measures (see the Appendix). The receivers were modeled as omnidirectional microphones suspended in space with no head between them. The head-shadow and pinna effects generated by use of a dummy head would have produced another spectral coloration, but, since such effects were all removed from the final stimuli (see the Appendix), there was no point in including them in the room model. Absorption coefficients for the internal surfaces of the room were all 0.3 for the reverberant room, giving a direct-to-reverberant ratio of −8.56 dB and −8.60 dB for the left-ear and right-ear impulse responses, respectively (high-pass filtered above 20 Hz). For the anechoic room, the coefficients were all set to 1, giving an infinite direct-to-reverberant ratio. Binaural stimuli were produced by generating the impulse responses for the two receivers in virtual space and convolving the speech samples with these two impulse responses.

F0 modulation of the target and the buzz was controlled orthogonally: (i) masker and target both monotonized, (ii) masker and target both modulated, (iii) masker monotonized and target modulated, (iv) masker modulated and target monotonized. Reverberation on the target and the masker was also controlled orthogonally: (i) masker and target both anechoic, (ii) masker and target both reverberant, (iii) masker anechoic and target reverberant, (iv) masker reverberant and target anechoic. The two experiments had therefore sixteen different conditions, covering two target modulations (0 versus ±2 semitones), two masker modulations, two target rooms (anechoic versus reverberant), and two masker rooms. ΔF0 was constant at two semitones. Each

of 160 target sentences was manipulated in four conditions (2 target modulations × 2 target rooms), creating 640 target stimuli. Four masking buzz stimuli were created (2 masker modulations × 2 masker rooms). All complex maskers and initial target stimuli (before changes in TMR by the adaptive procedure) were presented at a level of 69 dB SPL (see the Appendix).

C. Procedure

The experimental session began with two practice runs using unprocessed speech presented diotically and masked by diotic buzz, in order to familiarize listeners with the task. The following 16 runs measured speech reception thresholds (SRTs), one for each of the 16 experimental conditions. While sentence materials remained in the same order for all listeners, the pseudorandom order of the conditions was rotated for successive listeners. Thus, across a group of 16 listeners, a complete rotation of the conditions was achieved: all sentences contributed equally to each condition, and effects of order and materials were counterbalanced.

SRTs were measured using a one-up/one-down adaptative threshold method (Plomp and Mimpen, 1979). In this method, an individual measurement is made by presenting ten target sentences one after another, each one against the same masker. The TMR was initially very low (−32 dB) and in the initial phase, listeners had the opportunity to listen to the first sentence a number of times, each time with a 4-dB increase in TMR. When they believed that they could first hear about half the words of the target sentence, listeners were instructed to attempt to type a transcript of the first sentence. The correct transcript was then displayed on a computer monitor, with five key words in capitals, and the listener self-marked how many key words he or she got correct. Subsequent target sentences were presented only once and self-marked in a similar manner; the level of the target speech was decreased by 2 dB if the listener had correctly identified three or more of the five key words or else increased by 2 dB. Measurement of each SRT was taken as the mean TMR at the last eight trials.

D. Equipment

A computer monitor was visible outside the booth window for trial-by-trial feedback and a keyboard was inside for transcript responses. Signals were sampled at 20 kHz and 16 bits, digitally mixed, D/A converted by a 24-bit Edirol UA-20 sound card and amplified by a MTR HPA-2 Headphone Amplifier. They were presented binaurally to listeners over Sennheiser HD650 headphones in a single-walled IAC sound-attenuating booth within a sound-treated room.

III. EXPERIMENT 1. RANDOM-PHASE BUZZ MASKERS

A. Rationale

According to each theory, harmonic enhancement or harmonic cancellation, some predictions can be made. If the benefit of a two-semitones ΔF0 between a target male voice and a buzz masker was due to harmonic enhancement, then it should be disrupted primarily for a reverberant modulated target, to a smaller extent for an anechoic modulated target and it should be intact for a monotonized target (anechoic or reverberant), regardless of the masker conditions. If the benefit was due to harmonic cancellation, then it should be disrupted primarily for a reverberant modulated masker, to a smaller extent for an anechoic modulated masker and it should be intact for a monotonized masker (anechoic or reverberant), regardless of the target conditions. The first experiment tested these two predictions.

B. Results

Figure 2 presents the mean SRTs measured in experiment 1. A repeated-measures analysis of variance with four within-subject factors (target modulation × masker modulation × target room × masker room) was conducted in order to determine the influence of each factor on SRT. There was no main effect of the target modulation [$F(1,15) = 0.8$, $p > 0.05$]. There was a main effect of the masker modulation: mean SRTs were lower (i.e., better performance) when the masker was monotonized rather than modulated [$F(1,15) = 151.4$, $p < 0.0001$]. There was a main effect of the target room:

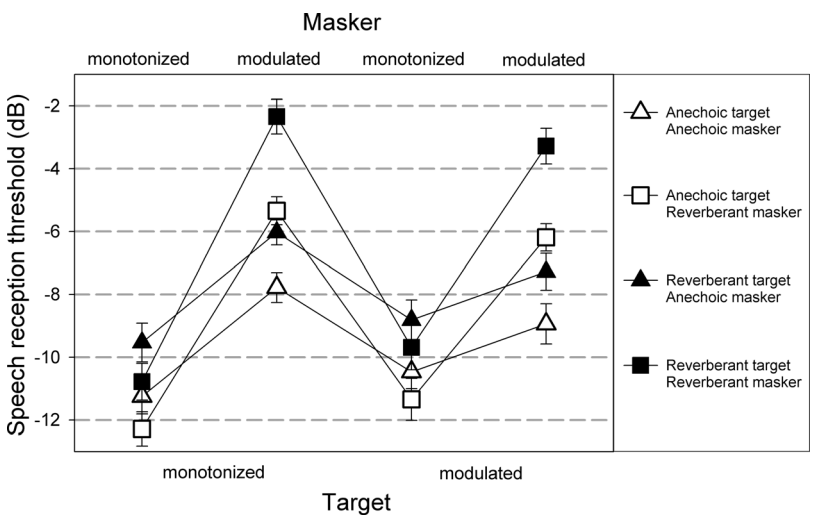

FIG. 2. Mean speech reception thresholds for the conditions where the target voice and the random-phase buzz masker were separated by a two-semitones ΔF0 and modulated factorially. Reverberation was also applied factorially to the target (empty versus filled symbols) and to the masker (triangles versus squares). Lower thresholds indicate greater intelligibility. Errors bars show ±1 standard error of the mean over the 16 listeners.

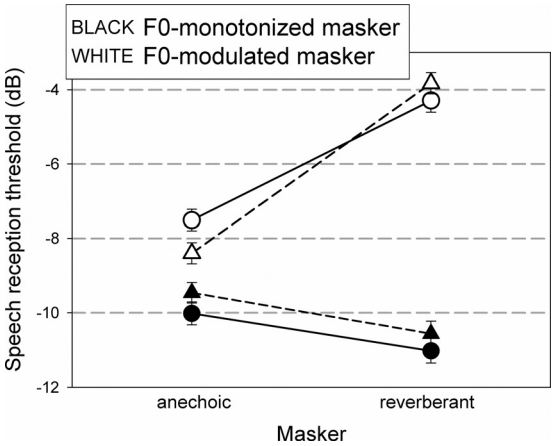

FIG. 3. Mean speech reception thresholds for the conditions where F0 modulation and reverberation were applied factorially to the buzz masker, averaged across all target configurations. The masker had random-phase partials in experiment 1 (circles) or sine-phase partials in experiment 2 (triangles).

mean SRTs were lower when the target was anechoic than reverberant [$F(1,15) = 99.8$, $p < 0.0001$]. There was also a main effect of the masker room: mean SRTs were lower when the masker was anechoic than reverberant [$F(1,15) = 26.1$, $p < 0.0001$]. Mean SRTs were averaged across target room and modulation (circles of Fig. 3) and across masker room and modulation (circles of Fig. 4) as a direct test of the predictions of harmonic cancellation and enhancement. As shown in Fig. 3, the masker room and masker modulation interacted strongly [$F(1,15) = 50.7$, $p < 0.0001$]. Target modulation and masker modulation also showed a modest interaction [$F(1,15) = 12.5$, $p < 0.01$], as illustrated in Fig. 5 (circles). No other interaction was significant.

C. Discussion

1. Harmonic cancellation

Figures 3 and 4 directly compared the predictions of harmonic cancellation and harmonic enhancement. In Fig. 3 (circles), mean SRTs were the lowest for the monotonized masker, increased by 2 or 3 dB for an anechoic F0-modulated masker and increased by 6 dB for a reverberant F0-modulated

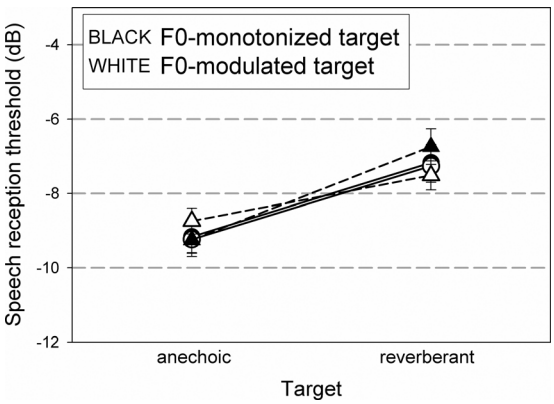

FIG. 4. Mean speech reception thresholds for the conditions where F0 modulation and reverberation were applied factorially to the target speech, averaged across all masker configurations. The masker had random-phase partials in experiment 1 (circles) or sine-phase partials in experiment 2 (triangles).

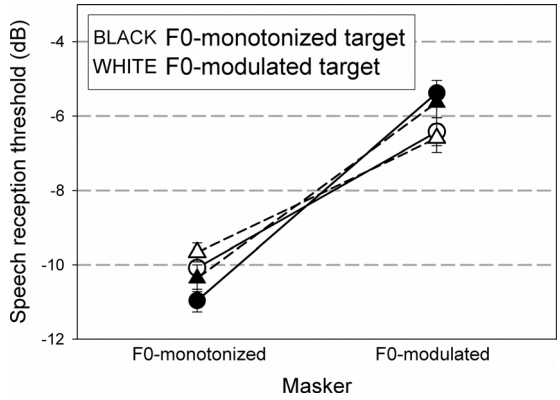

FIG. 5. Mean speech reception thresholds for the conditions where F0 modulation was applied factorially to the masker and to the target, averaged across all room configurations. The masker had random-phase partials in experiment 1 (circles) or sine-phase partials in experiment 2 (triangles).

masker. The results were fully consistent with the harmonic cancellation theory. Cancellation of the harmonic structure based on the masker's F0 is likely to underlie the benefit that listeners gain from a two-semitones ΔF0 between voice and buzz maskers. In anechoic conditions, with a 5-Hz rate of F0 modulation and a ±2-semitones width, the temporal resolution of the cancellation mechanism might be a little too sluggish to follow this rate of F0 modulation; the harmonicity of the buzz is blurred and the buzz cannot be cancelled as effectively as when it is monotonized, i.e. purely harmonic, resulting in a 2–3 dB elevation in SRTs. In reverberation, the F0 modulation provides the cancellation mechanism with many simultaneous F0s for the same buzz masker and therefore cancellation of an F0-modulated masker is further impaired under reverberation, resulting in a 6 dB elevation in SRTs. Note that in conditions where one source was F0-modulated while the other was monotonized, the competing F0s were alternately closer and further apart. However, SRTs were not elevated when the masker was monotonized and the target F0-modulated. So, the simple effect of F0 modulation of the masker was presumably not due to a fluctuation of instantaneous ΔF0s.

The harmonic enhancement theory predicted that loss of intelligibility should occur when the target became inharmonic, e.g., when the target was F0-modulated in reverberation. The data showed that this was not the case: There was no interaction between target room and modulation, as shown in Fig. 4 (circles). The meaning of the interaction, illustrated in Fig. 5 (circles), between the target modulation and masker modulation remains unclear. Somehow, an F0-modulated voice was easier to understand when the masker was itself F0-modulated, but harder to understand when the masker was monotonized. In any case, this was a weak interaction.

2. Degradation of target speech

In all conditions, intelligibility suffered when target speech was subject to reverberation, resulting in a 2-dB elevation of SRTs in the present data, as illustrated in Fig. 4. Degradation of target speech in reverberation should be expected independent from harmonicity effects; it can occur

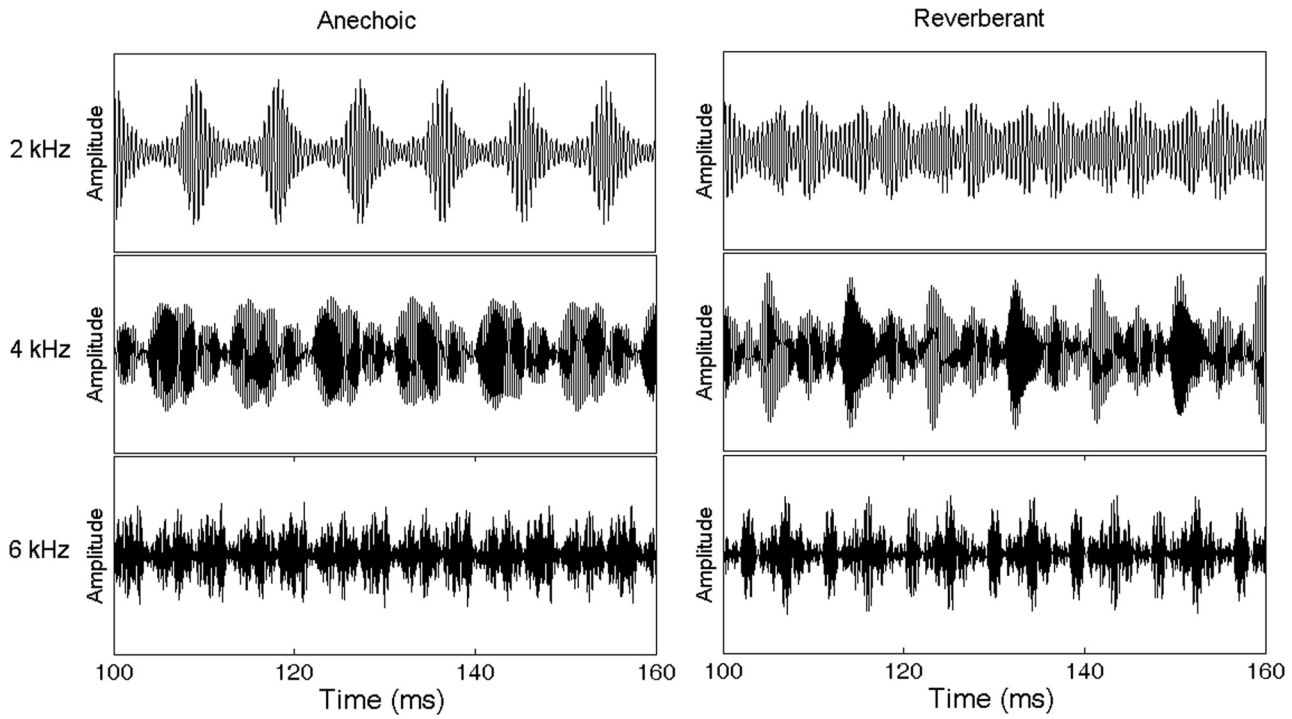

FIG. 6. Outputs of simulated auditory filters centered at 2 (top), 4 (middle) and 6 (bottom) kHz for the anechoic (left) and reverberant (right) monotonized random-phase buzz masker used in experiment 1. Amplitude is in arbitrary units, with equal scale for all signals.

even without any masker, and reflects the loss of amplitude modulation of the target due to reverberation. It is related to speech intelligibility indices, like the speech transmission index (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985).

IV. EXPERIMENT 2: SINE-PHASE BUZZ MASKERS

A. Rationale

As an alternative to harmonic cancellation, there is another possible explanation why harmonic sounds are less effective maskers than inharmonic sounds: Harmonic complexes might have within-channel temporal envelopes that are more modulated than inharmonic complexes. This envelope modulation could allow listeners to "listen in the dips" within each cycle of the fundamental period. In order to visualize these envelopes at different places along the basilar membrane (BM), the masker stimuli were passed through a simulation of rounded-exponential auditory filters with level dependency based on the data of Glasberg and Moore (1990), and with realistic phase responses based on the data of Oxenham and Dau (2001). Figure 6 shows the filtered waveforms at 2, 4, and 6 kHz for the anechoic (left panel) or reverberant (right panel) random-phase buzz masker used in experiment 1: the filtered envelopes are not strongly modulated. In contrast, Fig. 7 shows the filtered waveforms at the same three center frequencies for an anechoic (left panel) or reverberant (right panel) speech-shaped sine-phase complex. The masker envelopes are more strongly modulated than those of the random-phase complex, and the difference grows larger with increasing center frequency. In addition, listeners could benefit from the nonlinear amplification of the BM which amplifies the target signal at dips in these highly modulated envelopes, resulting in a better audibility of the signal than if compression had not occurred (Kohlrausch and Sander, 1995; Carlyon and Datta, 1997a; Summers and Leek, 1998). However, reverberation reduces dips in the masker envelopes (right panel of Fig. 7), so listeners would face a serious challenge with any type of reverberant maskers if they relied on dip-listening facilitated by BM compression. Using a speech-shaped sine-phase harmonic masker, experiment 2 replicated the design of experiment 1 to determine whether dip-listening could at least partly explained the benefit of masker's harmonicity in F0-segregation. If so, one would expect the differences between SRTs for modulated and unmodulated F0s and for anechoic and reverberant rooms to be larger with sine-phase maskers than they were with random-phase maskers.

B. Results

Figure 8 presents the mean SRTs measured in experiment 2. A repeated-measures analysis of variance with four within-subject factors (target modulation × masker modulation × target room × masker room) was conducted in order to determine the influence of each factor on SRT. There was no main effect of the target modulation [$F(1,15) = 0.2$, $p > 0.05$]. There was a main effect of the masker modulation: mean SRTs were lower when the masker was monotonized rather than modulated [$F(1,15) = 104.5$, $p < 0.0001$]. There was a main effect of the target room: mean SRTs were lower when the target was anechoic than reverberant [$F(1,15) = 57.4$, $p < 0.0001$]. There was also a main effect of the masker room: mean SRTs were also lower when the masker was anechoic than reverberant [$F(1,15) = 36.5$, $p < 0.0001$]. Mean SRTs were averaged across target room

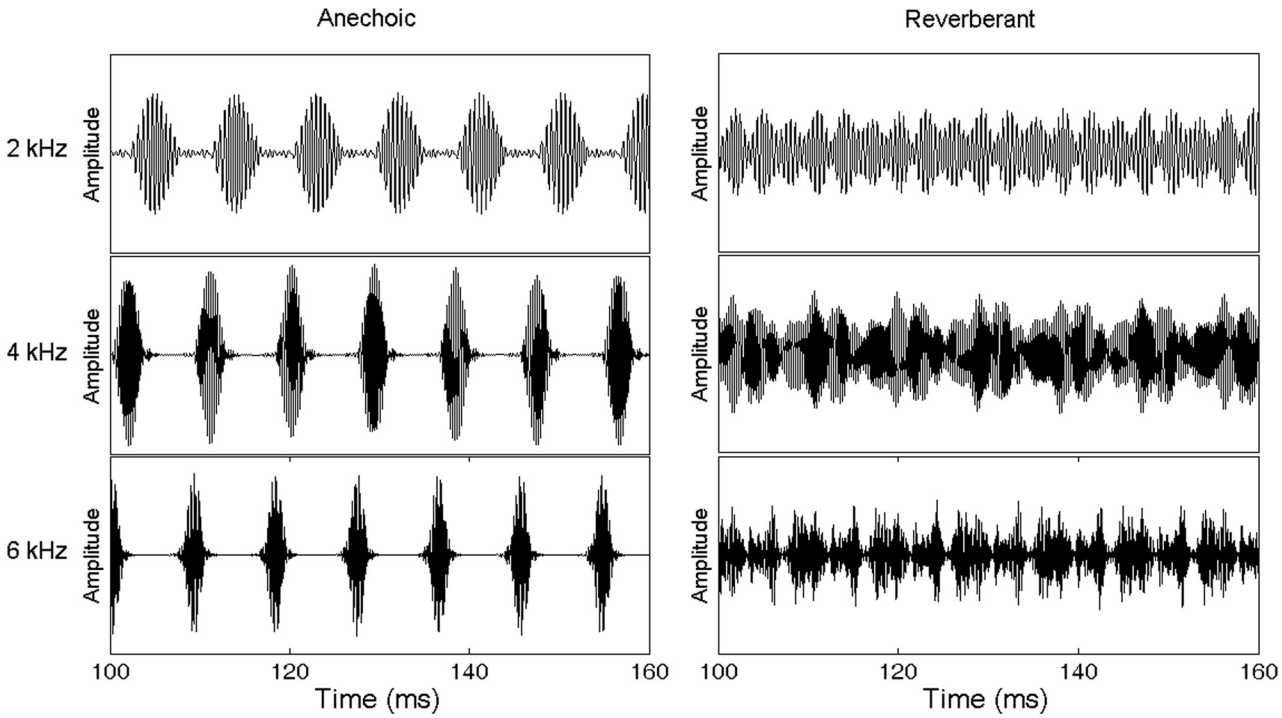

FIG. 7. Same as Fig. 6 but for a sine-phase buzz masker.

and modulation (triangles of Fig. 3) and across masker room and modulation (triangles of Fig. 4) as a direct test of the predictions of harmonic cancellation and enhancement. The masker room and masker modulation interacted strongly [$F(1,15) = 262.3$, $p < 0.0001$]. The target room and target modulation interacted slightly [$F(1,15) = 6.3$, $p < 0.05$]. Target modulation and masker modulation also showed a modest interaction [$F(1,15) = 12.9$, $p < 0.01$], illustrated in Fig. 5 (triangles). No other interaction was significant.

C. Discussion

1. Harmonic cancellation

In essence, the results were similar to those observed in experiment 1. In Fig. 3 (triangles), mean SRTs were the lowest for the monotonized masker, increased by 1 or 2 dB for an anechoic F0-modulated masker and increased by 6 dB for a reverberant F0-modulated masker. These results were again fully accounted for by the harmonic cancellation theory. In Fig. 4 (triangles), the data were contrary to the predictions of harmonic enhancement: SRTs were lower for an F0-modulated reverberant target than for a monotonized reverberant target. The meaning of this interaction as well as the interaction between F0 modulations of both sources remains unclear. In any case, those interactions were of small magnitude.

2. No role for dip-listening

Had dip-listening been involved in F0-segregation, one might have expected this mechanism to be seriously disrupted with a reverberant monotonized masker. In the right panels of Fig. 7, the phase randomizing effect of reverberation had

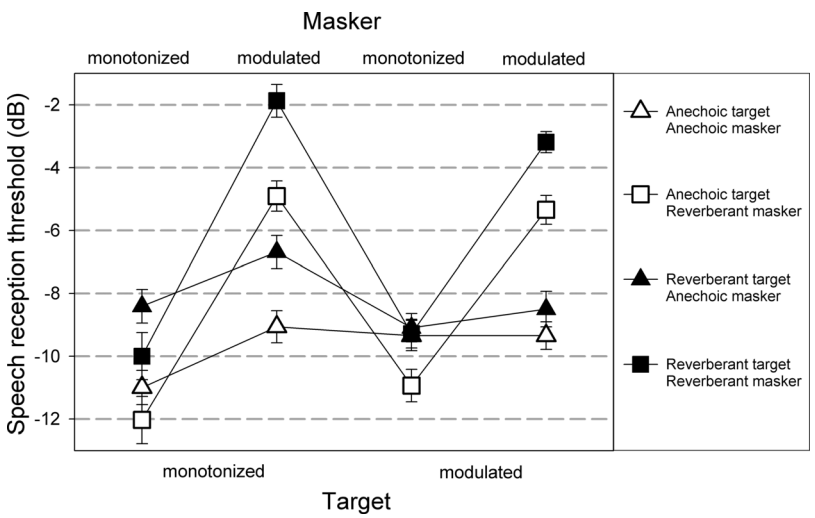

FIG. 8. Same as Fig. 2 but for a sine-phase buzz masker.

largely eliminated the dips in the masker envelopes at all center frequencies. Despite the flattening of the masker envelopes, SRT was lower when the monotonized buzz was reverberant than when it was anechoic. Therefore, the benefit of masker's harmonicity appeared to be independent of the listeners' ability to glimpse information about the target voice at dips in the period of the sine-phase buzz masker.

In order to further examine the effect of masker partials phase, a mixed factor analysis of variance with five factors (four within-subject factors as described in each experiment and one between-subject factor, the masker partials phase) was conducted. As expected, there were three main effects: masker modulation [$F(1,30) = 253.4$, $p < 0.0001$], target room [$F(1,30) = 147.9$, $p < 0.0001$] and masker room [$F(1,30) = 62.5$, $p < 0.0001$]. Target modulation and masker modulation interacted [$F(1,30) = 25.2$, $p < 0.0001$] and masker modulation and masker room interacted [$F(1,30) = 206.3$, $p < 0.0001$]. The masker partials phase only interacted with the two-way interaction of masker modulation and masker room [$F(1,30) = 4.4$, $p = 0.045$] as illustrated in Fig. 3. Among the four SRTs observed with a sine-phase masker, three of them were slightly shifted upward compared to the respective SRTs for a random-phase masker. So the meaning of this weak interaction seems to be that the detrimental effect of F0 modulation alone (i.e., in anechoic conditions) was smaller with a sine-phase masker than with a random-phase masker. This interaction was small compared to the main effects and did not support the major prediction of a role for dip-listening, that SRT for the anechoic monotonized masker would be lower when the masker had sine-phase partials than when it had random-phase partials and that the addition of reverberation would destroy that advantage.

V. GENERAL DISCUSSION

A. Contributions to the ΔF0 effect

Throughout the literature on ΔF0 effects, several underlying mechanisms have been proposed. With vowels as experimental stimuli, the effect occurs for very small values of ΔF0 (Scheffers, 1983). As a consequence, it is difficult to disentangle the relative contribution of harmonicity based mechanisms from that of the beating of close competing partials (Assmann and Summerfield, 1994; Culling and Darwin, 1994; de Cheveigné, 1999). With speech as experimental stimuli, the beating of competing partials is unlikely to play a role because the amplitude of the speech partials is constantly fluctuating. As a consequence, even when the target and masker were both monotonized, so that competing partials fall very close to each other along the BM (for instance, the eighth partial of the target's F0 and the ninth partial of the masker's F0), the resultant beating would be masked by the intrinsic modulations of speech. Note that this very-low-frequency beating produced by competing partials close in frequency is different from the beating produced by unresolved partials of the same complex and which was the object of the comparison between the two experiments (discussed further in Sec. V B).

On the other hand, with speech as experimental stimuli, the possible intrusion of informational masking makes it difficult to disentangle the relative contribution of harmonicity based mechanisms from that of streaming by F0 (Darwin *et al.*, 2003). The present study focused on the release from energetic masking provided by a ΔF0 between two competing harmonic sources. So the stimuli were chosen to avoid a role for attention and all possible cues of auditory grouping. The buzz maskers did not sound like speech at all. Although they had the same long-term excitation pattern as the speech stimuli, listeners were not confused as to which stimulus they ought to attend to. So it is very unlikely that informational masking was involved in the present segregation task and therefore unlikely that F0 was used as a perceptual grouping cue. Had streaming by F0 played a role, one might have expected speech and buzz to be most confusing to listeners when their F0s overlapped. However, Fig. 5 showed that the mean SRT for the condition where target and masker were both F0-modulated was lower than the mean SRT for the condition where only the masker was F0-modulated. So there was no indication in the data supporting the idea that streaming by F0 played any role.

Culling *et al.* (2003) attempted to extend to running speech the results observed with double-vowels by Culling *et al.* (1994). A masking talker differed from a target talker by about ten-semitones ΔF0 and a 15% shorter vocal tract, i.e., feminizing the masking voice. In their experiment 1, naturally intonated speech was more affected by reverberation than was monotonized speech. Two possible interpretations could explain such a result. First, reverberation might have affected F0-segregation by disrupting the harmonic structure of speech. The present results are in line with such an interpretation as cancellation of an F0-modulated masker is particularly difficult in reverberation. Second, reverberation might have affected processing of prosody, which monotonized speech lacks. The present results found strong impairments, despite the fact that both F0 manipulations removed meaningful prosody. In a second experiment, Culling *et al.* (2003) attempted to disentangle those two interpretations, by creating a third type of speech stimulus, in which the F0 pattern was inverted from the natural intonation. Such F0-inverted speech had as much variation of F0 as intonated speech, but was not expected to contribute to speech intelligibility. Their results showed that intonated speech was about equally affected by reverberation as F0-inverted speech and more affected than monotonized speech. Therefore both studies suggest that the detrimental effect of reverberation on intonated speech is related to disruption of harmonicity (particularly that of the masking voice) rather than disruption of prosody. Note that reverberation might potentially affect streaming by F0, but this ability occurs when competing voices compete for attention, a situation that the present experiments were designed to avoid.

Thus, the present results confirmed that in regard to energetic masking, a mechanism based on harmonicity of the masker is used to segregate a voice from speech-like harmonic maskers. We cannot be sure that harmonic enhancement does not play a role at higher TMRs. However, one may question how useful such a mechanism would be, since the target is very intelligible at positive TMRs. The auditory system is only challenged at negative TMRs. The present

data support the idea that when attempting to understand a voice at adverse TMRs of about −8 to −10 dB, listeners rely more on internally cancelling the harmonic structure of the maskers than internally enhancing the harmonic structure of a low-level voice.

B. Harmonicity versus phase effects

Detection of a pure tone in the presence of a harmonic complex masker is lower for masker partial phases giving a highly modulated waveform than for phases giving a less modulated waveform (Kohlrausch and Sander, 1995; Carlyon and Datta, 1997a; Summers and Leek, 1998). Furthermore, these phase effects in masking are strongly dependent on the masker level (Carlyon and Datta, 1997b; Summers and Leek, 1998). These results led to the idea that fast-acting compression of the BM could enhance the internal representation of a signal at dips within a masker period, thereby accounting for the poor masking ability of complexes with deep envelope modulations across auditory filters.

In an inharmonic complex, in which partial frequencies are jittered from their harmonic positions, envelopes are weakly modulated even within individual filters passing many partials, because partials beat at different rates than F0. Thus, one interpretation of the poor masking ability of harmonic complexes compared to inharmonic complexes might be that the BM amplifies greatly a target signal at dips in the deep envelope modulations of harmonic maskers but cannot enhance the representation of the same signal when masked by inharmonic maskers because their envelope modulations fluctuate less.

The present results do not support this interpretation for two reasons. First, the masker partial phases, random phase in experiment 1 or sine phase in experiment 2, did not materially influence the results, despite the fact that envelopes were more modulated with sine-phase than with random-phase buzz maskers (left panels of Figs. 6 and 7). Second, the phase-randomizing effect of reverberation, which eliminated dips in the masker envelopes across auditory filters, did not result in elevated SRTs as long as the masker remained monotonized. In conclusion, the present results showed not only that F0-segregation relies upon the masker's harmonicity but also that it is independent of the depth of masker envelope modulations across auditory filters. In other words, it is unlikely that a form of "listening in the dips" enhanced by the fast-acting compression of the BM could account for the poor masking ability of harmonic complexes observed here.

The fact that F0-segregation is not influenced greatly by the depth of within-channel envelope modulation suggests instead that F0-segregation is dominated by low-order harmonics. Culling and Darwin (1993) were interested in discovering which frequency region underlies the ΔF0 benefit. They synthesized vowels with an F0 in the region of the first formant peak, which was different from the F0 in the region of higher formant peaks. A ΔF0 in the first formant region largely accounted for the benefit. Bird and Darwin (1998) extended the results of Culling and Darwin (1993): they resynthesized speech sentences that were filtered into different bands above and below 800 Hz. Again, a ΔF0 in the frequency region below 800 Hz was necessary for the effect to occur. Interestingly, the auditory system is poorly sensitive to the phase of resolved harmonics (Moore and Glasberg, 1989) and autocorrelation of a pure tone disregards its starting phase. So it seems plausible that F0-segregation relies on the within-channel autocorrelation of resolved harmonics to extract the masker's periodicity. Such a mechanism would not only be insensitive to masker partial phase but also robust to the phase-jumbling effect of reverberation, as long as F0 remains steady. Indeed, the addition of two sinusoids at the same frequency but with different starting phases is just another sinusoid at that frequency. So, autocorrelation is robust to reverberation applied on resolved harmonics. In contrast, when F0 varies, autocorrelation would suffer from the multiplicity of periodicities within a channel.

VI. SUMMARY

The present experiments tested the theories of harmonic enhancement and cancellation as accounts for the beneficial effect of a two-semitones ΔF0 between speech and harmonic maskers. Harmonicity of the competing sources was disrupted by processes that could occur in realistic environments: F0 modulation and reverberation. The combination of these two factors resulted in large impairments when applied to the masker, but not when applied to the target. Thus, the ΔF0 effect seemed strongly dependent on the masker's harmonicity, not that of the target.

Interestingly, the masker partial phases, sine phase or random phase, did not influence the results. Moreover, no impairment was observed for a reverberant masker as long as it remained monotonized, while the phase randomizing effect of reverberation flattened the masker envelope modulations. Thus, the results did not support the notion that the benefit of masker's harmoncity could be accounted for by a form of "listening in the dips" of a modulated masker waveform.

The results are currently best explained by the mechanism of harmonic cancellation. When the masker is based on a different F0 than that of the target, here two semitones apart, listeners appear to internally suppress the masker harmonic structure and detect the target signal in the residue from this cancellation. When the harmonic structure of the masker is disrupted, it is not cancelled as effectively as a purely harmonic waveform and consequently masks the target speech more effectively.

ACKNOWLEDGMENTS

This work was supported by the UK EPSRC. We wish to thank Ray Meddis and an anonymous reviewer for their thoughtful comments on this manuscript.

APPENDIX

All source sentences (originally recorded by the MIT talker) were at the same rms level. The F0 manipulations, performed by the PRAAT PSOLA speech analysis and resynthesis package, introduced small variations in rms level. Table I shows these variations for different values of mean

TABLE I. Level of speech stimuli, in dB SPL, averaged over 80 sentences, after different F0 manipulations by the software PRAAT.

F0 re: 110 Hz (sem.)	Monotonized F0	Modulated F0 (± 1 sem.)	Modulated F0 (± 2 sem.)
0	68.48	68.48	68.47
1	68.67	68.67	68.65
2	68.86	68.85	68.84
3	69.03	69.03	69.01
4	69.19	69.18	69.16
5	69.31	69.31	69.31
6	69.44	69.45	69.44
7	69.57	69.57	69.56
8	69.69	69.69	69.68
9	69.79	69.79	69.78
10	69.88	69.87	69.87
11	69.96	69.96	69.96
12	70.05	70.05	70.04

F0: the higher the F0, the higher the rms level. In contrast, variations of the width of F0 modulation had no effect. To eliminate the small rms level difference that would occur between the complex maskers (110-Hz F0) and the target sentences (123.5-Hz F0), an initial rms equalization was performed by multiplying the signal amplitude by a correcting factor.

A further change in rms level was produced by the acoustic response of the reverberant room which amplified some frequencies and not others, producing a spectral coloration plotted on the top panel of Fig. 9. The middle and high frequency regions of the spectrum were affected by this spectral coloration. Since this frequency range contributes to speech intelligibility, it was necessary to equalize the spectra. We used a filter that compensated for the coloration produced by the reverberant room. The coloration being slightly different for left and right ears, we used two compensating filters, one for each ear. The excitation patterns of both the anechoic and reverberant sentences were used to create this compensating filter. We used the MATLAB-function FIR2 to design a finite impulse response (FIR) filter with 5000 coefficients, whose frequency response was the difference between the excitation patterns of the reverberant sentence and that of the respective anechoic sentence. We then applied this filter to the reverberant sentence and compensated the delay induced by convolution with the filter. The result of this equalization is illustrated in the bottom panel of Fig. 9 in which the final excitation patterns of anechoic and reverberant sentences are overlaid.

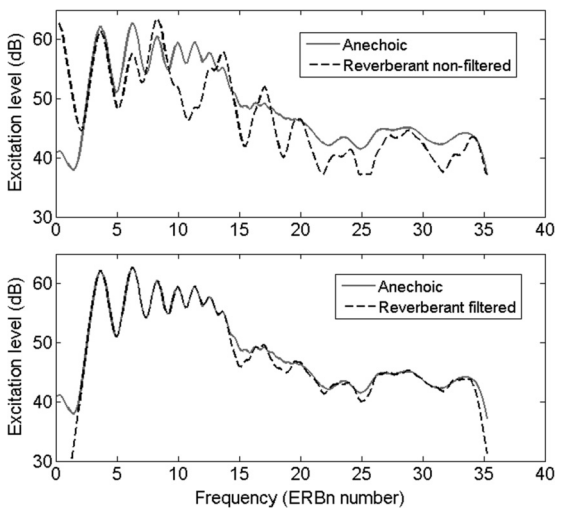

FIG. 9. Left-ear excitation patterns of a monotonized sentence after convolution with the anechoic or reverberant room impulse responses before filtering (top). Left-ear excitation patterns of the two convoluted signals, once the reverberant signal has been filtered to have the same excitation pattern as the anechoic one (bottom).

Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950.

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

Assmann, P. F., and Summerfield, Q. (**1994**). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. **95**, 471–484.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A.Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Brokx, J., and Nooteboom, S. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics. **10**, 23–36.

Carlyon, R. P., and Datta, A. J. (**1997a**). "Excitation produced by Schroeder-phase complexes: Evidence for fast-acting compression in the auditory system," J. Acoust. Soc. Am. **101**, 3636–3647.

Carlyon, R. P., and Datta, A. J. (**1997b**). "Masking period patterns of Schroeder-phase complexes: Effects of level, number of components, and phase of flanking components," J. Acoust. Soc. Am. **101**, 3648–3657.

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

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997**). "Concurrent vowel segregation. I. Effects of relative amplitude and F0 difference," J. Acoust. Soc. Am. **101**, 2839–2847.

de Cheveigné, A. (**1999**). "Waveform interactions and the segregation of concurrent vowels," J. Acoust. Soc. Am. **106**, 2959–2972.

Ciocca, V., and Darwin, C. J. (**1993**). "Effects of onset asynchrony on pitch perception: Adaptation or grouping?," J. Acoust. Soc. Am. **93**, 2870–2878.

Culling, J. F. (**1996**). "Signal processing software for teaching and research for psychoacoustics under UNIX and X windows," Behav. Res. Methods Instrum. Comput. **28**, 376–382.

Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. **93**, 3454–3467.

Culling, J. F., and Darwin, C. J. (**1994**). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.

Culling, J. F., Summerfield, Q., and Marshall, D. (**1994**). "Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels," Speech Commun. **14**, 71–95.

Culling, J. F., Hodder, K., and Toh, C. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.

Darwin, C. J., and Ciocca, V. (**1992**). "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component," J. Acoust. Soc. Am. **91**, 3381–3390.

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

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Houtgast, T., and Steeneken, H. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acout. Soc. Am. **77**, 1069–1077.
IEEE (**1969**). "IEEE recommended practise for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 227–246.
Kohlrausch, A., and Sander, A. (**1995**). "Phase effects in masking related to dispersion in the inner ear. II. Masking period patterns of short targets," J. Acoust. Soc. Am. **97**, 1817–1829.
Lea, A. (**1992**). "Auditory models of vowel perception," Doctoral dissertation, University of Nottingham, UK.
Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–134.
McKeown, J. D., and Patterson, R. D. (**1995**). "The time course of auditory segregation: Concurrent vowels that vary in duration," J. Acoust. Soc. Am. **98**, 1866–1877.
Meddis, R., and Hewitt, M. J. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233–245.
Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Moore, B. C. J., and Glasberg, B. R. (**1989**). "Difference limens for phase in normal and hearing-impaired subjects," J. Acoust. Soc. Am. **86**, 1351–1365.
Oxenham, A. J., and Dau, T. (**2001**). "Towards a measure of auditory-filter phase response," J. Acoust. Soc. Am. **110**, 3169–3178.
Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. **60**, 911–918.
Peterson, P. M. (**1986**). "Simulating the response of multiple to a single source in a reverberant room," J. Acoust. Soc. Am. **80**, 1527–1529.
Plomp, R., and Mimpen, A. M. (**1979**). "Speech-reception threshold for sentences as a function of age and noise level," J. Acoust. Soc. Am. **66**, 1333–1342.
Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, Rijksuniversiteit Groningen, The Netherlands.
Summerfield, Q., and Assmann, P. F. (**1991**). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony," J. Acoust. Soc. Am. **89**, 1364–1377.
Summerfield, Q., and Culling, J. (**1992**). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency," J. Acoust. Soc. Am. **92**, 2317(A).
Summers, V., and Leek, M. R. (**1998**). "Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners," Hear. Res. **118**, 139–150.
Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.