Levels in Speech Communication: Relations and Interactions
C. Sorin et al. (Editors)
© 1995 Elsevier Science B.V. All rights reserved.

Perceiving vowels in the presence of another sound: a quantitative test of the "Old-plus-New" heuristic*

C. J. Darwin

Experimental Psychology, University of Sussex, Brighton BN1 9QG, U.K.

1. ABSTRACT

When two sounds overlap in time, can a listener subtract perceptually one sound from the other? The experiment described here provides a quantitative test of listeners' ability to subtract a 500-Hz tone from a following vowel. /ɪ/-/ɛ/ phoneme boundaries along a first-formant (F1) continuum were measured for 56-ms vowel continua in which the energy of the 500-Hz component had been increased by various amounts up to +9.54 dB from its original level. As expected, the phoneme boundary moved to lower nominal F1 frequencies as the energy of the 500-Hz component was increased. These phoneme boundaries were compared with those obtained from continua which contained an additional 500-Hz tone. The tone started 100 ms before the vowel and stopped either at vowel offset or 100 ms after it. It was presented at different levels depending on the gain of the 500-Hz component in the vowel. The level of the additional tone was chosen to give a constant +9.54 dB overall gain at 500 Hz during the vowel. For all but the highest amplitude of the additional tone, listeners were able to use a difference in onset time alone to subtract perceptually the amplitude of the preceding tone from the overall amplitude at that frequency during the vowel. An additional difference in offset time of 100 ms produced an extra perceptual subtraction of around 2-3 dB regardless of level.

2. INTRODUCTION

We normally listen to speech against a background of other sounds. The presence of these other sounds raises a problem that is common both to the perception of speech and to the perception of sound in general. How can parameters that describe properties of a single sound source be extracted from a signal which may also contain other sounds?

An extreme example is silence. The silence that cues the closure period of a voiceless stop is not absolute, it is the silence of a particular speaker. But except in the ideal conditions of a sound-proof room, it will not appear as absolute silence in the information passed to the brain by the ear. There will generally be other sounds present in the listener's environment at that time. One must presume that the brain represents a voiceless stop as including silence, since it could not foresee the infinite variety of other sounds that might be present during any particular utterance of a voiceless stop. But the signal from the ear to the brain during any particular voiceless stop will be anything but silent. How does the brain infer the presence of relative silence?

* DSP programming and equipment were provided by grant GR/F 45356 from the Computational Science Initiative of the S.E.R.C., and other equipment and general computing facilities were provided by S.E.R.C. grants GR/G 27430 and GR/H 29001. Robert Hukin supported by M.R.C. grant G9116710N helped in the running of the experiment. Valter Ciocca and Brian Roberts generously commented on a draft and Quentin Summerfield made helpful reviewer's comments.

One general class of solution proposes, following Bregman (1990), that there are low-level primitive grouping mechanisms that provide a preliminary sorting of sound elements into putative sound sources. These proposed low-level processes exploit simple properties of sounds such as harmonicity or common onset-time, in order to segregate different groups of sound elements.

If such segregation is to work efficiently, the brain must be able to subtract from the total sound present, those elements that correspond to one putative source, leaving those elements that belong to the remaining sound sources. Such a subtraction mechanism can be seen clearly at work in numerous demonstrations of auditory streaming where a single tone sequence breaks into separate streams, destroying the melodic contour and rhythm of the original single stream. Here the sound elements that are being segregated are well-separated in frequency and in time. But it is less obvious that such perceptual subtraction can occur when different sound sources overlap in frequency and in time - as is generally the case with the complex sounds of our normal environment.

With such complex sounds, the pattern of activity at the ear (the auditory excitation pattern) will be the sum of all the sounds that are simultaneously present. The appropriate perception of a particular sound would be facilitated if the brain were able to subtract from this composite auditory excitation pattern the contribution made by other sounds.

Four types of experiments that have varied the onset time of different sounds have provided direct evidence from a variety of paradigms for some such perceptual subtraction.

First, when one harmonic of a complex starts earlier than the rest, the complex is judged to be purer than when the harmonics are all synchronous (Bregman and Pinker, 1978). If the complex is a vowel, and one of the harmonics in the first formant region starts earlier than the others, the vowel's phonetic quality changes in a way that is compatible with the asynchronous harmonic making a reduced contribution to the vowel (Darwin, 1984). Furthermore, if energy is added to a vowel either at a harmonic frequency (Roberts and Moore, 1991) or at a different frequency (Roberts and Moore, 1991) its effect on the vowel's quality is reduced when it starts before the vowel.

These experiments on vowel quality rely on subjects being able to judge the *relative* intensity of simultaneous frequency components. Such an ability is also needed in the second, 'profile analysis' task where listeners must detect an increment in the level of one component of a complex sound despite a random change of overall level between the two observation intervals. The threshold increment in level in this task increases substantially if the component carrying the increment starts earlier than the other components (Green and Dai, 1992). One explanation for this result is that the difference in onset time cues listeners to subtract the asynchronous tone from the remainder. Such subtraction then yields different perceptual entities - the asynchronous tone as one entity and the remaining components as a second. Intensity comparisons across different objects are then assumed to be harder than those within a single object.

A third type of experiment that shows similar effects of onset time is in pitch perception. A mistuned component that starts earlier than the rest of an otherwise harmonic complex makes less contribution to the pitch of the complex than if it had started at the same time (Darwin and Ciocca, 1992). This result can be explained by assuming that the asynchronous mistuned component is perceptually subtracted from the rest of the complex for the purpose of calculating the pitch of the complex.

A fourth type of experiment involves the perceptual consequence of the illusory continuity of one sound when another alternates with it. If the two sounds are spectrally identical, the less intense of the alternating sounds may be heard as continuous with an additional sound pulsing against its background (Warren *et al.*, 1972). There is some evidence that the energy of the less intense (perceptually continuous) sound has been subtracted from the energy in each frequency channel of the more intense (pulsating sound) to give it a reduced loudness. Warren (1982, p. 141) reports that when 80 dB and 83 dB noise bursts are alternated, the pulsating sound has the same loudness as the (80 dB) illusorily continuous sound, but that when their intensities are 80 and 82 dB the pulsating sound is quieter than the continuous. These observations have been extended recently(Warren *et al.*, 1994) confirming that some perceptual subtraction occurs both for tones and for noise. The apparently continuous sound appears to be subtracted from the more intense pulsating sound to give it a reduced loudness compared with when it is presented alone. Moreover, this reduction in loudness is greater as the quieter (apparently continuous sound) is made more intense.

Fifth, there are a number of phenomena described as "auditory after-effects" or as due to "auditory enhancement" that are formally similar to some of the above experiments. For example, Summerfield and his collaborators have demonstrated that either a broad band noise or a harmonic complex with a flat-spectrum can be identified as a vowel if it is preceded by a similar sound that has spectral valleys at the frequencies of the perceived vowel's formants (Summerfield *et al.*, 1984; Summerfield and Assmann, 1987). Neither the precursor sound alone, nor the flat-spectrum test sound can be identified as a vowel. Deeper spectral valleys (5 dB) give better identification of the vowel than do shallower (2dB). Again this result could be due to the perceptual subtraction of the earlier from the later sound.

Experimental results such as these are compatible with the brain using what Bregman has called the Old-plus-New heuristic:

> "if part of a present sound can be interpreted as being a continuation of an earlier sound, then it should be" (Bregman, 1990, p222).

Bregman's definition of the heuristic begs the quantitative question of how much is perceptually subtracted. Is the earlier sound assumed to continue at the same level, or, can it be assumed to increase in level and so take over *all* the energy at its frequencies? A second question concerns the nature of the subtraction: in what units does the subtraction occur? The subtraction might occur in amplitude (units linear with pressure) or energy (units proportional to the square of pressure) or in some other perhaps neurally-based unit. Warren's data for tones is better matched by subtraction of amplitude than by subtraction of energy, but his data on noise is better matched by subtraction of energy.

The interpretation of the heuristic that is compatible with Warren's results and that is favoured by Bregman assumes that the earlier sound continues into the present sound at the same loudness. Any increase in level at the frequency of the earlier sound that occurs as the present sound starts is then attributed to the present sound. The effect of this heuristic is shown in the left panel of Figure 1 applied to a complex tone with an additional tone starting earlier at the frequency of one of its harmonics. The additional tone is subtracted out, leaving the original complex. A quantitative implication of this interpretation is that a less-intense earlier sound will cause less energy to be subtracted from the present sound leaving more attributed to the present sound. The experiments described above are all compatible with this first interpretation of the heuristic, which we will call Level-Dependent Subtraction. Level-Dependent Subtraction could be (though need not be) explained by a relatively peripheral mechanism such as adaptation in the auditory nerve.

A second, alternative interpretation assumes that *all* the energy in the present sound that is at the frequency of the earlier sound is heard as a continuation of the earlier sound. We will call this interpretation Level-Independent Subtraction. A quantitative implication of this interpretation is that reducing the intensity of the earlier sound will have no effect on the amount of that sound that is perceptually subtracted from the present sound and consequently will not influence the amount attributed to the present sound. None of the experiments described above critically rules

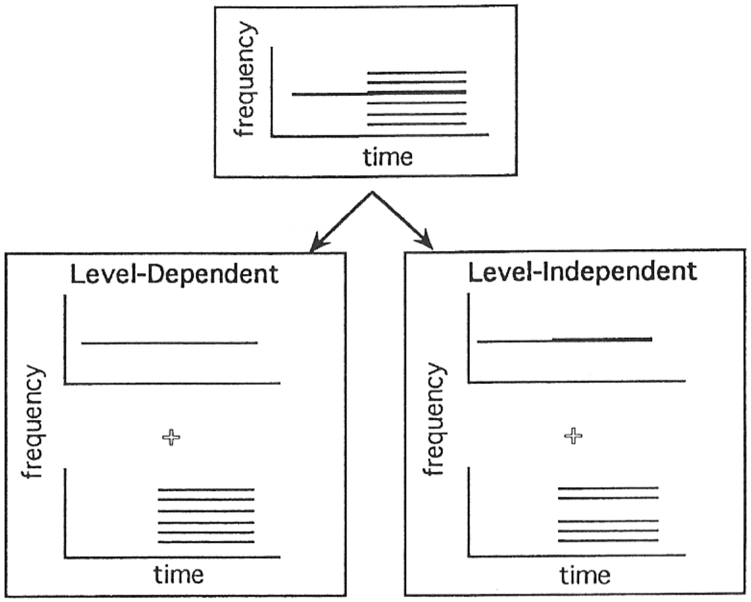

Figure 1. Stylised spectrogram of a composite sound consisting of a complex tone to which an additional sound at the frequency of the fourth harmonic has been added. The additional sound starts before the complex, and during the complex increases the total energy at its frequency. Level-Independent Subtraction would segregate the composite into one tone that contained all the energy at the frequency of the additional tone, plus the original complex tone minus the component at that frequency. Level-Dependent Subtraction would lead to a segmentation that only removes some of the energy at the frequency of the additional tone from the complex.

out the second alternative of Level-Independent Subtraction. The mechanism underlying such effects could not be adaptation, but could be a simple attentional process that allocates different peripheral auditory channels to different sound sources.

There are three recent findings that are compatible with the second interpretation. These experiments have all shown results where a reduction in intensity of a leading sound does not influence the effect produced by it.

First, Carlyon (1989) measured the threshold for a brief 1 kHz tone masked by a simultaneous bandpass noise with a notch at the frequency of the target. The masked threshold for such a tone is decreased when the target and mask are preceded by another sound (the primer) which consists of either one or both bands of the notched-noise masker. The striking finding that Carlyon reported was that this decrease in threshold was independent of the level of the primer over a 30 dB range. This independence of level invalidates explanations of the priming effect based on mechanisms such as adaptation or perceptual subtraction of absolute levels in each frequency channel.

Second, using a similar profile analysis task to that of Green and Dai (1992) described earlier, Dai (personal communication) has demonstrated that the increase in threshold produced by a leading tone is only reduced by 3 dB when the leading tone is decreased in intensity by 20 dB.

Third, Ciocca and Darwin (in preparation) using a similar paradigm to their earlier work described above have shown that a 20 dB decrease in the intensity of a mistuned leading tone does not alter the reduction in pitch shift caused by its earlier onset-time.

The results of these three experiments all show that a substantial reduction in the intensity of a leading component has rather little effect on the perceptual changes that it has induced. As such, they are more compatible with the second Level-Independent interpretation of the Old-plus-New heuristic. (Of course, there must be *some* level-dependent effect even in these experiments at least around detection threshold for the leading component.)

The purpose of the following experiment was to seek direct evidence to test the first, Level-Dependant interpretation of the Old-plus-New heuristic. It asks whether the effect of a preceding harmonic on vowel quality depends on the level of the preceding tone, or whether it is independent of level.

The vowel task that is used can estimate accurately, if indirectly, the perceived intensity of a particular component. The phoneme boundaries that it measures are directly related to the perceived first formant frequency which in turn are directly related to the relative intensities of the frequency components making up the vowel. Perceptually-induced loudness changes in a particular frequency component can be calibrated against *actual* changes in level, by equating shifts in the vowel phoneme boundary. The calibration against actual reductions in level also allows us to ask whether the units of the subtraction are amplitude or level or neither?

The experiment investigates in addition whether the effect of asynchrony between a tone and a vowel is attributable entirely to adaptation by examining whether an additional difference in *offset* time enhanced the effect of onset time.

3. EXPERIMENT

In this experiment we use changes in the phoneme boundary between /ɪ/ and /ɛ/ along a first-formant continuum to estimate the perceived level of a 500-Hz harmonic close in frequency to the first formant. The rationale for the experiment is as follows. If we simply boost the energy at 500 Hz in each member of the continuum (by adding +6 dB of energy at 500 Hz in phase to raise the total level at 500 Hz by 9.54 dB) there will be a shift in the phoneme boundary compared with the original continuum. If we then gradually reduce this energy back to the original level (by adding to the original harmonic progressively less extra energy), there will be a gradual return of the phoneme boundary back to its original value. We then have a calibration curve for phoneme boundary values against the actual reduction in level from the maximum 9.54 dB boost. We can then use this calibration to examine the *perceptual* subtraction of another added tone from the vowel.

This added, leading tone (also at 500 Hz) starts before the (possibly boosted) vowel and either ends with the vowel or after it. It has a different level depending on the boost that has already been given to the 500-Hz component of the vowel. The leading tone's level is sufficient to raise the *total* energy at 500 Hz during the vowel back to +9.54 dB. For example, if the leading tone is being added to a vowel from the original continuum, then it will have to have an energy of 6 dB greater than the existing 500-Hz component of the vowel in order to raise the total energy to +9.54 dB. But if it is being added to a vowel, whose 500-Hz component has already been increased by 6 dB, then the leading tone need only be the same amplitude (+0 dB) as the 500-Hz component in the original (unboosted) continuum to give a total boost of +9.54 dB.

The result of these manipulations is that we end up with a set of conditions in which, during the vowel, the 500-Hz component is always boosted by 9.54 dB, but where different levels of tone at 500 Hz protrude from this vowel, either forwards in time or forwards and backwards. If the listener were able to subtract the amplitude of the protruding tone from the total amplitude at that frequency during the vowel, then we should get the same phoneme boundary shift as if the actual energy during the vowel had been reduced by that amount. We can then compare the phoneme boundary shifts obtained with different levels of protruding tones, with our calibration curve. If listeners show level-dependant subtraction based on subtraction of amplitude, then their phoneme boundaries with the leading tones added should correspond to those in the corresponding condition without the leading tone. If the listeners show level-independent subtraction, then the phoneme boundaries should not change with the different levels of leading tone.

3.1. Stimuli and Method

The experimental stimuli consisted of 16 vowel continua, each of 7 members. These continua were all derived from an original vowel continuum that varied in first formant frequency (F1) to give a percept that changed from /I/ at lower F1 values to /ε/ at higher values.

The original continuum had 7 members that varied in first formant frequency from 396 Hz to 521 Hz in steps of 21 Hz. The second and third formants were fixed at 2100 and 2900 Hz. The bandwidths of the first three formants were 90, 110 and 170 Hz. The fundamental frequency was 125 Hz, and the total duration was 56ms (16 ms rise/fall plus 24 ms steady state).

From this basic continuum, three groups of five experimental continua were derived. The 6 continua of the first group ('no tone') had the level of the 500Hz component boosted by 0, +3.97, +6.0, +8.44, +9.17 and +9.54 dB. All levels are given relative to the level of the 500 Hz tone in the corresponding member of the original continuum.

In the 5 continua of the second group ('onset'), the level of the 500 Hz component of the vowel was similarly boosted by 0, +3.97, +6.0, +8.44 and +9.17 dB relative to its level in the corresponding member of the original continuum, but a 500-Hz leading tone at various levels was also added; it started 100 ms before the vowel but stopped at the same time as the vowel. The levels of the 500Hz tone were +6, +3, 0, -9 or -18 dB relative to the level of the 500-Hz component of the corresponding member of the original continuum. These levels were chosen so that the *total* energy at 500 Hz during the vowel was always +9.54 dB regardless of the level of the additional 500-Hz leading tone. The tone's starting phase was such that it always added in phase with the 500-Hz component of the vowel.

The 5 continua of the third group ('on-offset') were similar to the second, but the 500-Hz tone started 100-ms before and stopped 100-ms after the vowel.

Stylised power spectra for one sound from the 'onset +6 dB' continuum (upper panels) and one from the 'onset -18 dB' continuum (lower panels) are shown in Figure 2. The left-hand panels show the spectrum of the leading tone; the right-hand panels show the spectrum of the vowel, with the energy of the 500 Hz tone divided into two parts. The upper dashed line indicates the energy corresponding to the tone; and the solid line the remaining energy. If the listener can perceptually remove the continuation of the leading tone from the vowel, then the solid line corresponds to the energy that should then be attributed to the vowel. These levels are also the *actual* energy levels used in the 'no tone' group of continua. Corresponding pairs of conditions are shown in Table 1.

In summary, the levels of the 500 Hz components in the various conditions were chosen so that if listeners perceptually subtract the amplitude of the leading tone from the vowel, then their phoneme boundaries should be identical in corresponding continua shown in Table 1. For example, when the leading tone is at a level of +6dB in the 'onset' continuum its phoneme boundary should be at the same position as the +0dB member of the 'no tone' continuum.

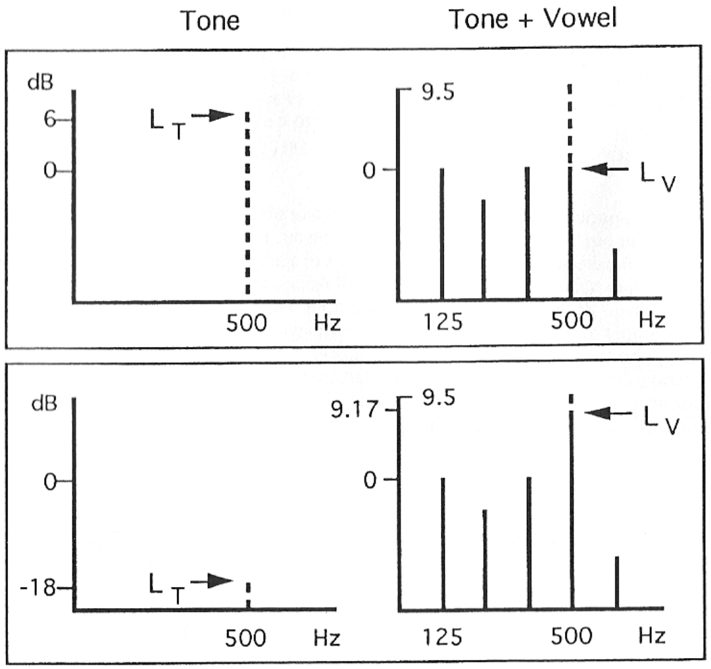

Figure 2. Stylised power spectra for one sound from the 'onset +6 dB' continuum (upper panels) and the corresponding one from the 'onset -18 dB' continuum (lower panels). The left-hand panels show the spectrum of the leading tone; the right-hand panels show the spectrum of the vowel, with the energy of the 500 Hz tone divided into two parts. The upper dashed line indicates the energy corresponding to the tone; and the solid line the remaining energy.

If listeners do *not* perceptually subtract the leading tone from the vowel, but rather regard all the energy at 500 Hz that is present during the vowel as being relevant to the vowel, then all the boundaries in the 'onset' and 'on-offset' groups of continua should be the same as for the +9.54 dB continuum in the 'no tone' group. The reason for this prediction is that the total energy at 500 Hz for all the continua in these groups is always +9.54 dB above the level of the corresponding member of the original continuum.

The sounds were synthesised in real time at 44.1kHz by a general-purpose synthesis program (Russell, 1992) using the 56001 processor of the Digidesign Audiomedia board on a Mac IIci which controlled the experiment. The vowel spectra were computed by adding harmonics with the amplitudes and phases given by the vocal tract transfer function equations in Klatt (1980). The sounds were output through the Audiomedia's 16-bit DACs and anti-aliasing filters, into passive attenuators and then into a custom-built headphone amplifier and Sennheiser 414 headphones. Subjects heard the sounds binaurally in a double-walled IAC booth. The level of the

500-Hz component of the member of the original continuum with an F1 of 500 Hz was 60 dB SPL.

The experiment was run in two blocks: one block contained the 'no tone' and the 'onset' groups of continua, the other the 'no tone' and the 'on-offset' groups. The 770 trials (10 replications of 7 sounds from each of 11 continua) of each block were presented in different pseudo-random orders in a session lasting about 20 minutes. Different subjects took the two blocks in different orders. Eight native speakers of British English with no history of hearing disorder took the experiment.

	Continuum condition				
	+ 0	+ 3.97	+ 6.0	+ 8.44	+ 9.17
500Hz component of vowel L_V	+ 0	+ 3.97	+ 6.0	+ 8.44	+ 9.17
500 Hz leading tone L_T	+ 6	+ 3.0	+ 0	- 9.0	-18.0
Total energy at 500Hz	+ 9.54	+ 9.54	+ 9.54	+ 9.54	+ 9.54

Table 1. Levels of 500 Hz components (added in phase) relative to the level of the 500 Hz component in the appropriate member of the original continuum.

3.2. Results

The numbers of /I/ responses made to each stimulus, averaged across the 9 subjects, are shown in Figure 3. Individual identification functions were somewhat steeper than the average since subjects have their phoneme boundaries in different places. The identification functions are well-behaved.

Individual phoneme boundaries were estimated from the proportion of 'i' responses given by each subject to the 10 replications of each 7-member continuum. The rescaled tanh function:

$$\frac{1}{1 + e^{-s(a-x)}}, \qquad \text{where a is the boundary and s the slope parameter}$$

provided a sufficiently good fit and allowed convenient boundary estimation.*

Following previous practice we refer to the first formant frequency that was used to synthesise a sound as its *nominal* F1. When we add energy to a harmonic of a vowel its nominal F1 stays constant by definition. Phoneme boundaries for the various continua are expressed in terms of this nominal F1. So, if addition of energy has no perceptual effect on the quality of a vowel (since, for example, a precursor has allowed the additional energy to be perceptually removed) then the nominal F1 boundary will be unchanged. However, if the additional energy produces the percept of a *higher* F1 frequency, then the phoneme boundary will appear at a *lower* nominal F1 value along the continuum.

The mean phoneme boundaries across the 9 subjects for each continuum are shown in Figure 4. For the 'no tone' conditions, the abscissa indicates the physical level of the 500 Hz component relative to the original continuum. As expected, the phoneme boundaries in the 'no tone' conditions (solid symbols) move systematically to lower nominal F1 values as the level of the 500 Hz component is increased. This movement indicates that as energy is added to the 500 Hz component, so the perceived F1 increases in frequency, causing the nominal F1 boundary value to decrease in frequency. These boundary values from a physical change in energy at 500 Hz provide a calibration for estimating the perceived level of the 500 Hz component in the onset and on-offset conditions. We can use them to measure how much energy has been perceptually subtracted from the physically present level for different levels of the onset or on-offset tones.

* Zoltan Dienes suggested the tanh function to me as an adequate analytic approximation to the cumulative normal distribution.

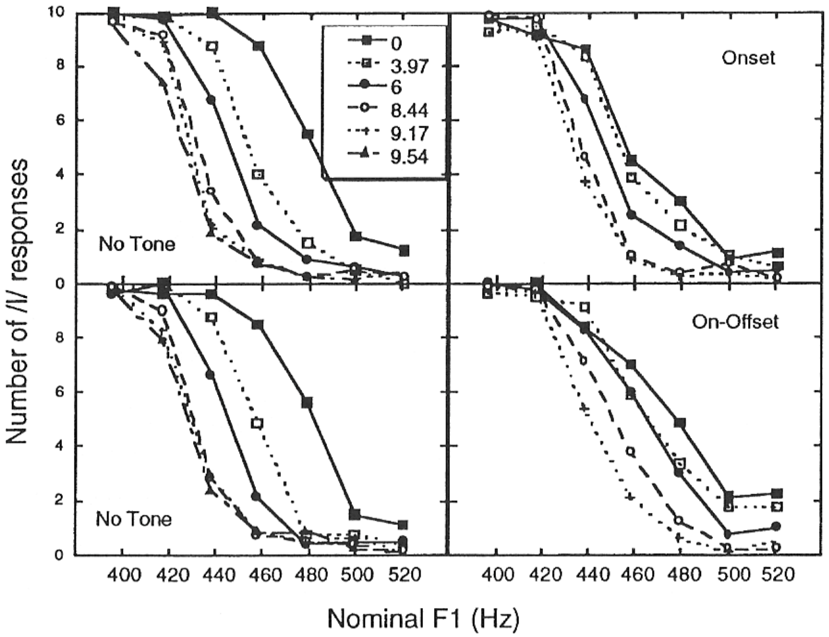

Figure 3. Identification functions for /I/-/ε/ continua. The average number of /I/ responses across 8 subjects are shown as a function of the nominal F1 frequency. The left hand panels show conditions where the 500-Hz component was physically increased in intensity by the number of dB shown in the legend. The right-hand panels show conditions where a leading tone was added to these conditions, increasing the energy of the 500-Hz component during the vowel to +9.54 dB. The tone either started earlier (upper panel) or both started earlier and stopped later than the vowel (lower panel). If subjects can perceptually subtract the amplitude of the leading tone from the vowel, the boundaries in the corresponding left- and right-hand panels should be identical.

For the onset and on-offset conditions (open symbols), the abscissa indicates the gain of the 500 Hz component of the vowel (L_V) that would result if perfect subtraction by amplitude occurred. The numbers in the figure against each set of data points give the level (L_T)of the onset and on-offset tones themselves. The reason for plotting the results this way is that if perfect subtraction were taking place, then the boundaries for the no tone conditions should be identical to those for the onset and on-offset conditions. On the other hand, if the perceived level of the 500 Hz component of the vowel were independent of the level of the precursor tone, then the phoneme boundaries in the 'onset' and 'on-offset' conditions should give a horizontal line in Figure 4.

The results show that the perceived level of the 500 Hz component of the vowel *is* influenced by the level of the onset or the on-offset tones. For low levels of the onset or on-offset tone, phoneme boundaries parallel those produced by the appropriate physical changes in the level of the 500 Hz component of the vowel, indicating some perceptual subtraction. At the highest (+6 dB) level of tone the subtractive effect begins to saturate.

Overall, boundaries in the 'on-offset' condition (open circles) are at higher F1 frequencies than in the 'onset' condition (open squares). This difference does not vary with level. The extent of this difference can be estimated in dB from Figure 3 by comparing both onset and on-offset conditions with the no tone conditions that show similar F1 boundaries. Including an offset tone at 500 Hz in addition to the onset tone is equivalent to subtraction of roughly an extra 2 to 3 dB from the 500-Hz component of the vowel.

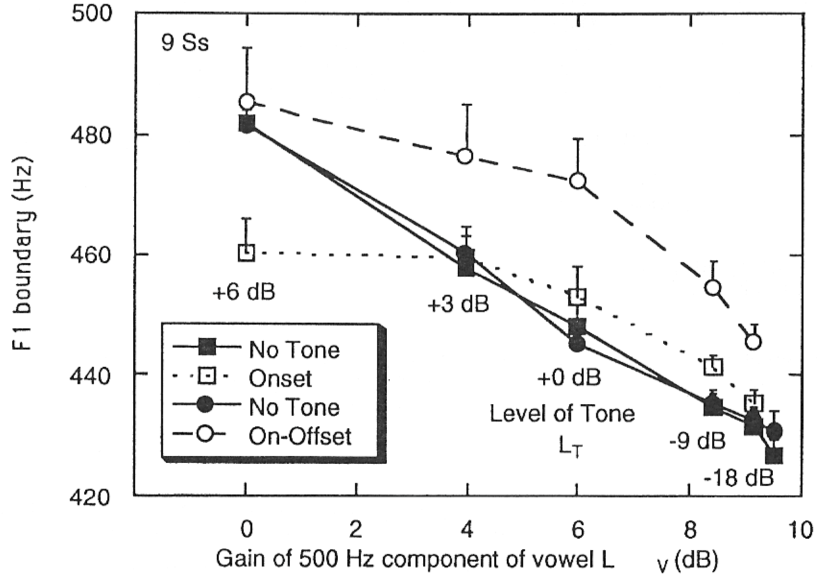

Figure 4. Nominal F1 phoneme boundaries (with standard error bars across 9 listeners) averaged across 8 subjects for /I/-/ε/ continua. The filled symbols show boundaries when the 500 Hz component has been physically increased in level by the amount (L_V) indicated on the abscissa. The open circles show boundaries for vowels where similar vowels contain an additional leading tone starting 100 ms before the vowel. The level of the tone (L_T indicated beside each data point) increases the energy of the 500 Hz component during the vowel to 9.54 dB above L_V=0. The abscissa for the open symbols gives the gain of the 500 Hz component without the extra tone. The open squares show boundaries when the leading tone starts 100 ms before and stops 100 ms after the vowel. If listeners can perceptually subtract the amplitude of the leading tone from the vowel, then boundaries for corresponding filled and open symbols should be identical.

3.3. Discussion

The main result of the experiment is that some form of perceptual subtraction is occurring in the perception of vowel quality. The perceived level of the physically constant 500 Hz component of the vowel changes systematically with the level of a 500 Hz tone that precedes (or precedes and follows) the vowel. This result is consistent with the Level-Dependent interpretation of the Old-plus-New heuristic described in the Introduction, and is inconsistent with the Level-Independent interpretation.

In the onset condition, at all but the highest level of preceding tone, the auditory system appears to subtract accurately the amplitude of a preceding 500-Hz tone from its subsequent amplitude in the vowel. The level resulting from this subtractive process is then used to estimate the F1 frequency. This subtractive process could be based on peripheral adaptation, since the duration of the onset tone (100 ms) is long enough for most of the adaptation found in the auditory nerve to have occurred (Smith, 1979). At least some of this effect must be of central origin though, since Darwin and Sutherland(1984) showed that the effect of an additional asynchronous tone on the /I/-/ε/ phoneme boundary could be partly reversed by causing the leading portion of the tone to group with another, harmonically-related tone that started with it, but stopped as the vowel started. Adaptation processes would not be influenced by this manipulation.

The subtractive mechanism revealed in this experiment appears to be based on subtracting amplitude rather level. The onset data of Figure 4 show a very good fit with the appropriate no tone condition, with the matching being based on the assumption that amplitudes are being subtracted. If the data were replotted assuming that levels rather than amplitudes were being subtracted, the onset and on-offset data would lie on a much steeper curve which would provide a very poor fit to the no tone condition. There is thus a discrepancy between the apparent subtraction of amplitudes in this experiment and the subtraction of levels suggested by Warren's (Warren, 1982) anecdotal report. The discrepancy may lie in the different paradigms used, or, more likely, it may lie in the fact that this experiment used tones whereas Warren's used noise. The recent results reported by Warren *et al.* (1994) are broadly compatible with the perceptual subtraction being carried out in terms of amplitude for tones, and in terms of energy for noise (although that issue is not addressed directly by the paper).

Letting the tone continue after the vowel, as well as starting before it, produces an additional subtraction of around 2 or 3 dB to that produced by a difference in onset time alone. This result confirms the previous finding by Darwin and Sutherland(1984) that offset times can influence the perceptual organisation of speech sounds. The effect of offset times cannot be due to adaptation but could be due to a more central grouping mechanism.

The results of experiments such as the one reported in this chapter, where a change in a phoneme boundary occurs as a result of perceptual grouping, argue against the notion that speech processing, by "pre-empting" the available auditory information, is immune from the results of auditory grouping (Remez *et al.*, 1994). As discussed previously (Darwin, 1991), it is difficult to see why the speech perception mechanism would choose to solve the problem of auditory sound segregation *de novo* rather than exploiting the accumulated wisdom of millions of years of pre-speech evolution.

At present, it is not clear why some paradigms that have looked at the influence of a preceding sound find effects that are independent of the level of that sound, whereas others find effects that do depend on their level. It is quite clear from the present experiment that the auditory system is capable of Level-Dependent subtraction of a precursor sound. Further experiments are needed to clarify the conditions under which it does not choose, or is unable to use the results of such subtraction.

REFERENCES

Bregman, A. S. (**1990**). *Auditory Scene Analysis: the perceptual organisation of sound.* (Cambridge, Mass, Bradford Books, MIT Press).

Bregman, A. S. and Pinker, S. (**1978**). "Auditory streaming and the building of timbre," Canad. J. Psychol. **32**, 19-31.

Carlyon, R. P. (**1989**). "Changes in the masked thresholds of brief tones produced by prior bursts of noise," Hear. Res. **41**, 223-236.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound: constraints on formant perception.," J. Acoust. Soc. Am. **76**, 1636-1647.

Darwin, C. J. (**1991**). "The relationship between speech perception and the perception of other sounds.," in *Modularity and the motor theory of speech perception* edited by I. G. Mattingly and M. G. Studdert-Kennedy (Erlbaum., Hillsvale,N.J.),pp. 239-259.

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. and Sutherland, N. S. (**1984**). "Grouping frequency components of vowels: when is a harmonic not a harmonic?," Quart. J. Exp. Psychol. **36A**, 193-208.

Green, D. M. and Dai, H. (**1992**). "Temporal relations in profile comparisons," in *Auditory physiology and perception* edited by Y. Cazals, L. Demany and K. Horner (Pergamon Press, Oxford),pp. 471-478.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971-995.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S. and Lang, J. M. (**1994**). "On the perceptual organization of speech," Psych. Rev. **101**, 129-156.

Roberts, B. and Moore, B. C. J. (**1991**). "The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony.," J. Acoust. Soc. Am. **89**, 2922-2932.

Russell, P. (**1992**). "*Synthesiser: User Manual*". Brighton, UK.: Laboratory of Experimental Psychology, University of Sussex

Smith, R. L. (**1979**). "Adaptation, saturation, and physiological masking in single auditory nerve fibers," J. Acoust. Soc. Am. **65**, 166-178.

Summerfield, A. Q. and Assmann, P. (**1987**). "Auditory enhancement in speech perception.," in *The psychophysics of speech perception* edited by M.E.H. Schouten (Martinus Nijhoff, Dordrecht),pp. 140-150.

Summerfield, A. Q., Haggard, M. P., Foster, J. and Gray, S. (**1984**). "Perceiving vowels from uniform spectra: phonetic exploration of an auditory after-effect.," Percepn. Psychophys. **35**, 203-213.

Warren, R. M. (**1982**). *Auditory Perception: a new synthesis* (New York, Pergamon).

Warren, R. M., Bashford, J. A., Healey, E. W. and Brubaker, B. S. (**1994**). "Auditory induction: reciprocal changes in alternating sounds," Percepn. Psychophys. **55**, 313-322.

Warren, R. M., Obusek, C. J. and Ackroff, J. M. (**1972**). "Auditory induction: perceptual synthesis of absent sounds," Science **176**, 1149-1151.

Levels in Speech Communication: Relations and Interactions
C. Sorin et al. (Editors)

On the contribution of instance-specific characteristics to speech perception

A. R. Bradlow, L. C. Nygaard, and D. B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405, U.S.A.

1. INTRODUCTION AND BACKGROUND

The role of variability in the listener's interpretation of the speech signal has been the topic of extensive research, and in general, it has been treated as a source of "noise" to be separated from the meaningful, abstract, symbolic units of speech [1,2]. For example, the general approach of many studies of speech acoustics has been to perform various measurements of speech sounds as produced various talkers in various phonetic and/or prosodic environments, e.g. [3-5]. The data are then used to derive generalizations about the nature of speech sounds and their contextual variation, which can then be used to investigate the acoustic cues to the related perceptual contrasts. An explicit assumption of this approach is that the variability inherent in the speech signal presents an "obstacle" to the listener that needs to be removed, or "stripped away", from the signal to facilitate perception of the underlying abstract linguistic units. Accordingly, the driving force behind this general research agenda has been the specification of the principles that underlie the observed variability in the speech signal so that it can be perceptually "recovered" by the listener.

In contrast, our theoretical approach treats variability of the speech signal as a useful source of information that is available to listeners at all stages in their interpretation of the speech signal [6-8]. For example, this approach predicts that listeners will be sensitive to inter-talker differences; and that, rather than removing this source of variability from the signal as a consequence of perceptual analysis, listeners use this information as a basis for identifying talker characteristics that can aid in the interpretation of the linguistic message. Accordingly, in our acoustic analyses of sentences produced by multiple talkers we have deliberately avoided averaging across many talkers to derive summary generalizations about speech production; rather, we focus on inter-talker differences and try to correlate these differences with differences in listener responses. In general, our approach contrasts markedly with the traditional, "abstractionist approach" to speech because we focus on instance-specific variation, as opposed to the traditional emphasis on instance-independent generalizations about idealized, abstract symbolic forms [9,10].

In keeping with this general theoretical orientation, the research presented in