A neural cancellation model of F_0-guided sound separation.

Alain de Cheveigné
alain@linguist.jussieu.fr
Laboratoire de Linguistique Formelle, CNRS/Université Paris 7,
2 place Jussieu, 75251, Paris, France.

ABSTRACT

Listeners were presented with pairs of concurrent vowels and requested to report one or two vowels. The ΔF_0 was either 0 or 6%, and RMS levels before mixing were either the same or different by 10 or 20 dB. Responses for each vowel within a stimulus were classified according to relative level (-20, -10, 0, 10, 20 dB) and ΔF_0(0 and 6%). Identification was better at ΔF_0=6%, and this effect was greatest when the target was weak (-20 and -10 dB). This outcome is difficult to account for with current models, but can be explained by invoking a within-channel neural cancellation filter. A model of concurrent vowel identification based on this filter is consistent with our experimental data, and agrees with results that show that the auditory system segregates harmonic sounds by cancelling the harmonic background.

1 INTRODUCTION

The harmonic structure of voiced speech plays an important role in our ability to understand speech in a speech background. Identification is degraded when that cue cannot be used, for example when target and interference voices have the same fundamental frequency (F_0) [14, 2, 25, 1, 11, 19]. In this paper we present experimental and modeling efforts that lead to somewhat counterintuitive conclusions: a) The harmonic nature of a sound does not protect it from interference. Instead, the harmonic nature of the interference makes it easier to ignore. b) The selectivity of the basilar membrane may play a secondary role in segregation of sounds based on harmonicity: a temporal model of segregation based on filtering *within* auditory channels accounts for our data better than a spectro-temporal model that chooses among channels separated by peripheral filtering.

Improvements in identification when an F_0 difference (ΔF_0) is introduced are usually attributed to a segregation mechanism that exploits the harmonic structure of voiced speech, and fails when target and background have the same F_0. A question that arises is whether the harmonic structure of the target, that of the interference, or both, determine segregation [5]. Summerfield and Culling [26] found that masked thresholds for vowels were lower when the masker was harmonic rather than inharmonic, and that target harmonicity made no significant difference. Lea [18] found better identification when the background vowel was voiced rather than whispered, but the voicing state of the target vowel had no significant effect. de Cheveigné et al. [7] found better identification for harmonic than inharmonic backgrounds, but also, paradoxically, that identification was better for inharmonic rather than for harmonic targets. A later experiment yeilded no effect of target [9]. Overall, there is little evidence that the auditory system enhances harmonic targets, and strong evidence that it cancels harmonic interference. Other results are congruent with these findings. Zissmann et al. [28] presented subjects with co-channel speech in which the masker could be attenuated when either the target or the background was voiced. Intelligibility was better in the latter case. Harmonic cancellation was also more effective to reduce co-channel speech interference in a speech recognition system [6], possibly because natural speech is not sufficiently stationary for harmonic enhancement to be effective [5].

The lack of evidence for harmonic enhancement is disturbing, as it implies that the harmonic structure of voiced speech does not protect it from interference. Harmonicity of interference is useful, but this doesn't solve the problem of speech in non-harmonic noise. Many models and methods exploit target harmonicity to enhance or group target components [10, 4, 17, 15, 16], and it plays an important role in Auditory Scene Analysis [3]. *Stimulus* harmonicity, rather than *target* harmonicity, might have the important effect of signaling a single source. Classic double-vowel experiments are blind to to this aspect because they force listeners to report two vowels for every trial. We use a modified task in which a subject may report one or two vowels.

F0-guided segregation has been explained by *spectral* models inspired from Parsons' work [24], and *spectro-temporal* models in the vein of Weintraub's system [27]. They differ in the degree of selectivity required of peripheral analysis: sufficient to resolve individual components for the former, sufficient to isolate spectral zones reflecting each vowel for the latter. Assmann and Summerfield [1] argued that peripheral selectivity is insufficient to support a purely spectral model. Meddis and Hewitt [22] proposed a successful spectro-temporal model that partitions the population of peripheral filter channels on the basis of the periodicity of their response. Segregation occurs between but not within channels, and thus its success depends on whether peripheral filtering excludes each vowel from at least some channels, from which the other vowel may be salvaged. In a third class, *temporal* models, segregation occurs within channels and peripheral selectivity is therefore less critical [5]. We develop such a model based on a hypothetical neural cancellation filter that removes certain spikes from spike trains carried by the auditory nerve.

2 EXPERIMENT

2.1 Methods

Six japanese subjects were presented with stimuli consisting of either a single vowel or the sum of two different vowels. For each stimulus the subjects had to answer either one or two vowel names. They thus simultaneously judged the number of vowels present within each stimulus, and identified them. The vowels were steady-state synthetic tokens of five Japanese vowels (/a/, /e/, /i/, /o/, /u/) produced at two F_0s (125 Hz and 132.5 Hz), with equal RMS signal level. Double vowel stimuli were formed by scaling one of the vowel waveforms by a factor (-20, -10, 0, 10 or 20 dB), adding them, and scaling the sum to an RMS level that was the same for all stimuli. The stimuli were presented to subjects via headphones at a sound pressure level between 63 and 70 dBA. Stimuli were 200 ms in duration with 20 ms raised-cosine onset and offset ramps. The stimulus set for each session comprised 600 double vowels and 240 single vowels, and each subject performed 5 sessions.

Single vowel stimuli were scored once. The vowel was deemed identified if its name appeared among the vowel(s) reported. Double vowels were scored twice, each vowel in turn being considered a target and the other vowel a background (interference). Identification rates were calculated as a function of the nature of the target, that of the interference, and their mutual relation (ΔF_0, relative level). In addition, the average number of vowels reported per stimulus was recorded for each condition.

The experiment differed from classic double vowel experiments in four ways: a) the stimulus set included single vowels, b) subjects were allowed to answer one or two vowels, c) the relative level between vowels in a pair was varied rather than fixed, d) the scoring method measured 'target-correct' rates instead of 'both-correct' rates.

2.2 Results

The number of vowels reported per stimulus is plotted in Fig 1. At unison, subjects reported two vowels quite often when both constituents had the same level, no doubt because the composite spectrum did not resemble that of any single vowel. When either consituent was stronger, they reported one vowel more often. Single vowels evoked double responses on about 10% of all trials (triangle to the right), but that proportion varied widely between subjects (2% to 27%). The target-correct identification rate is plotted in Fig 2. Identification was better at ΔF_0=6% than at unison, especially when the target was weak (-20 or -10 dB) relative to the interfering vowel. At higher levels the ΔF_0 effect was small.

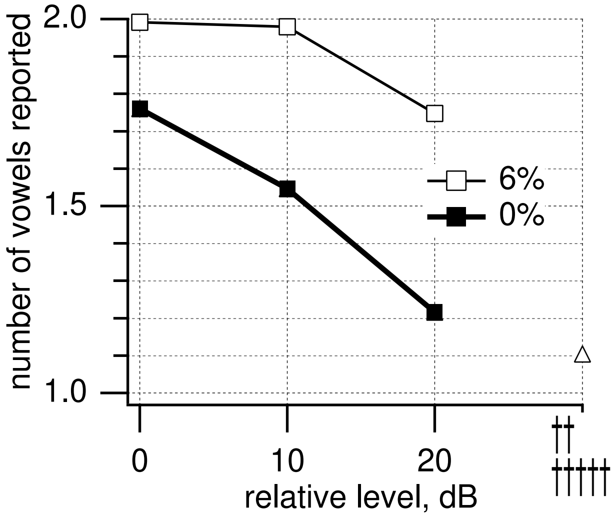

Fig. 1 *Number of vowels reported per stimulus as a function of relative level between vowels, at unison (filled symbols) and ΔF_0 = 6% (open symbols). Triangle at right is for single vowels.*

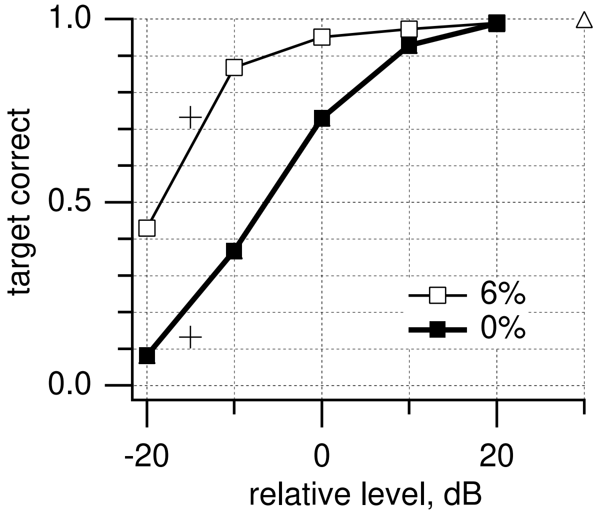

Fig. 2 *Identification rate as a function of the level of the target relative to the interfering vowel, at unison (filled symbols) and ΔF_0=6% (open symbols). Triangle at right is for single vowels. Crosses represent data obtained at -15 dB in another experiment with the same subjects.*

When the target is weak, estimation of its F_0 should be difficult while that of the interference should be easy. Large ΔF_0 effects for weak targets thus support the hypothesis of harmonic cancellation, at the expense of that of harmonic enhancement. The latter mechanism might be effective when targets are strong, but any benefit it brings is masked by the ceiling effects. The pattern of results varied somewhat between vowel pairs. Fig 3 shows results for one particular pair (/o/+/u/). We note that for both vowels the effect of ΔF_0 is strong when the vowel is at -20 dB. This result will be used to compare our model to that of Meddis and Hewitt [22].

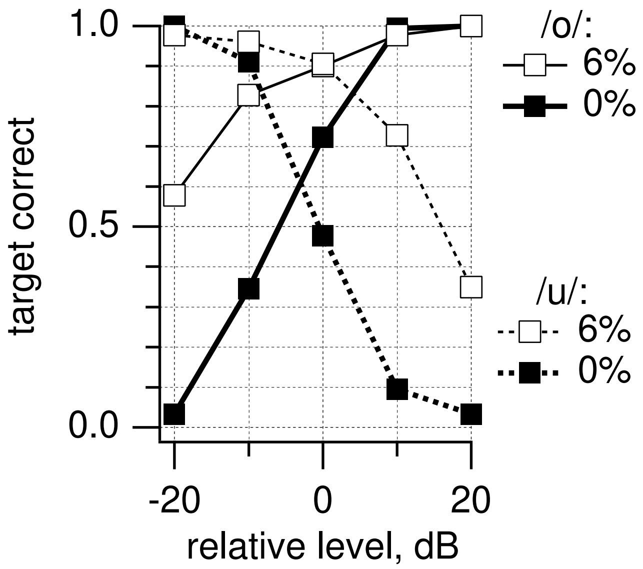

Fig. 3 *Identification rate of /o/ (ascending lines) and /u/ (descending lines) within an /o/+/u/ pair, as a function of the level of /o/ relative to /u/, at unison (filled symbols) and ΔF_0=6% (open symbols).*

3 MODELS

3.1 Meddis and Hewitt's model

The model of Meddis and Hewitt [22] exploits the harmonic structure of only one vowel, the dominant one. When the target is weak, the background is dominant, and the model thus performs cancellation. In this respect it is compatible with our results. Meddis and Hewitt showed that their spectro-temporal model was superior to that of Assmann and Summerfield [1], which in turn had been shown to be superior to purely spectral (place) models. Palmer [23] tested it with physiological data recorded in the 8th nerve of the guinea pig in response to double vowels, and found it plausible. The model comprises a stage of peripheral filtering, followed by hair cell transduction, followed by the calculation of an autocorrelation function (ACF) within each channel. ACFs of all channels are added to obtain a summary function (SACF) and the largest peak of this function is used to estimate the dominant period. This part of the model is similar to the same authors' pitch perception model [21]. The periodicity of each channel is then examined, and a partition is made between channels that have a peak at the dominant period and those that do not. SACFs are calculated for both groups, and the short-lag portion (below 4.5 ms) of each SACF is used to identify the two vowels present. At unison, all channels respond with the same period and so a partition is impossible, which explains why identification is less good than when there is a ΔF_0.

Essential is the assumption that peripheral filtering provides some channels that are not completely dominated by the stronger vowel. If all channels are dominated by the same vowel, no partition is possible, and the model predicts no ΔF_0 effect. The situation might arise when one vowel is stronger than the other, as in our experiment. Meddis and Hewitt's model was applied to our stimuli. At intermediate levels (-10 to 10 dB) the partition occured as expected, but at -20 or 20 dB some pairs showed no partition, as illustrated by Fig. 4 for the pair /o/+/u/. The proportion of channels dominated by /o/ increases with relative level, and at 20 dB it has reached 100%. The model therefore predicts no ΔF_0 effect for /u/ at that level. Our experimental results show a clear ΔF_0 for the same condition (Fig. 3), that the model thus cannot explain. To be fair, a different choice of filter shape (deeper skirts) or of partition criterion (less stringent) might allow the model to work over a wider range. Nevertheless it is worth investigating whether segregation might be performed *within* channels, in which case characteristics of peripheral filtering would be of less importance.

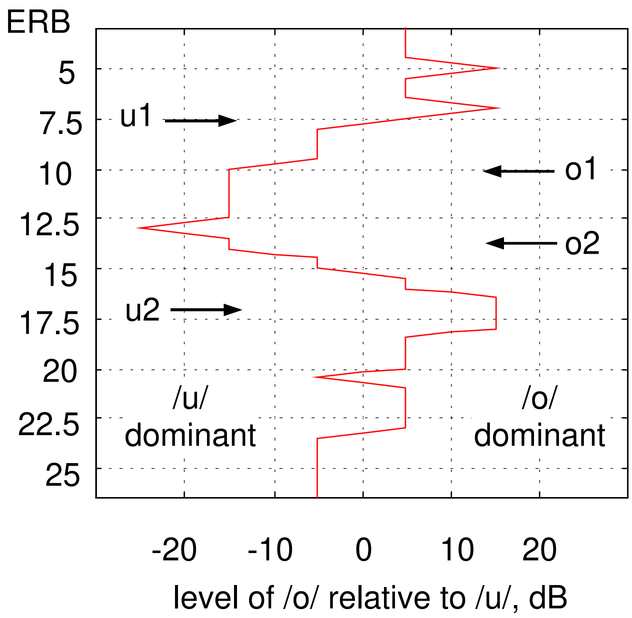

Fig. 4 *Periodicity that dominates channels as a function of relative level in Meddis and Hewitt's model. To the left of the crooked line the dominant modulation is 125 Hz (/u/). To the right it is 132.5 Hz (/o/). Arrows indicate the first two formants of each vowel.*

3.2 Neural harmonic cancellation filter

Suppose that each channel (group of fibers of similar characteristics) is processed by a neuron that is driven via two pathways, one direct and excitatory, and the other delayed and inhibitory (Fig. 5). Suppose further that every spike that travels along the direct path is transmitted *unless* a spike arrives, within a certain time window, along the delayed path. The filter will thus weed out all intervals of duration equal to the delay from the spike train. It turns out that this is sufficient to suppress the correlates of one vowel if the filter is tuned to that vowel's period [5].

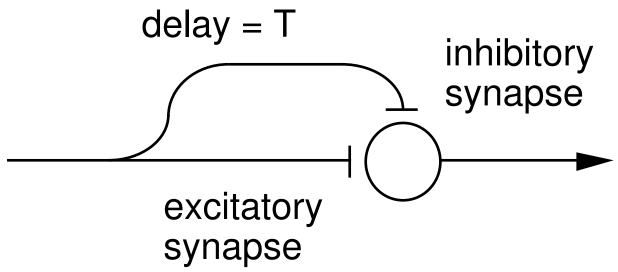

Fig. 5 *Neural harmonic cancellation filter.*

The effect of the filter may be crudely approximated by:

$$o(t) = MAX(0, i(t) - i(t - T))$$

where i(t) and o(t) are respectively the discharge probability at the input and output of the filter and T is the delay. The MAX operation reflects the fact that probability cannot be negative. The filter was applied to each output channel of a gammatone-filterbank/hair-cell model [20, 13] with input consisting of the stimuli used in the experiment. All filters were tuned to a common delay equal to the period of one vowel, and the outputs were analyzed for evidence of the other vowel. The array of filtered discharge probabilities may be interpreted and exploited in several ways, two of which are illustrated below.

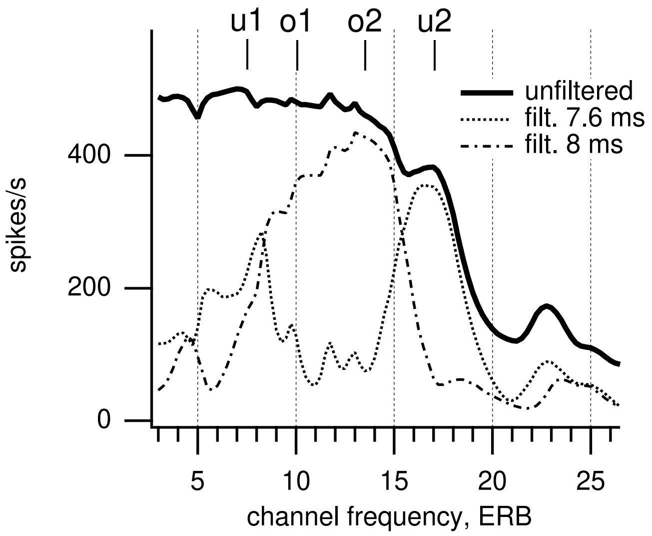

Fig. 6 *Discharge probability as a function of channel frequency on an ERB scale. Thick line is before the cancellation filter, thin lines are after cancellation of /o/ (dots) or /u/ (dots & dashes).*

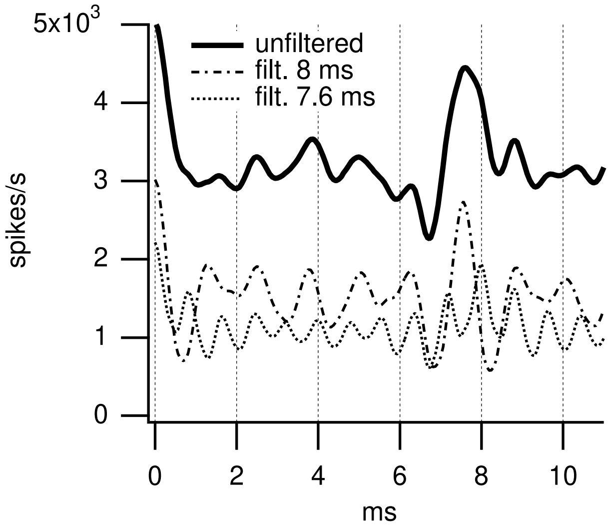

Fig. 7 *Square root summary autocorrelation function. Thick line is before the cancellation filter, thin lines are after cancellation of /o/ (dots) or /u/ (dots & dashes).*

3.2.1 Place

Fig. 6 represents discharge probability as a function of channel frequency (on an Equivalent Rectangular Bandwidth scale) in response to the double vowel /o/+/u/ at 0 dB relative level and ΔF_0=6%. The thick continous line is the output of the hair cell model. The thin dotted lines represent the output of the cancellation filter when it is is tuned to the period of either vowel. When the filter is tuned to the period of /o/ (7.6 ms), peaks stand out at the first two formants of /u/. When it is tuned to the period of /u/, a 'hump' seems to reflect the first two formants of /o/.

3.2.2 Time

Instead of rate vs place, the cancellation filter output may be interpreted as a time domain pattern, for example by calculating the summary autocorrelation function. Fig. 7 shows the square root of the SACF in response to the same stimulus as for Fig. 6 (the square root compensates for the quadratic nature of autocorrelation and improves intelligibility of the plot). When the filter is tuned to the period of /o/ (7.6 ms), the SACF has a shape characteristic of /u/ and a peak at its period (8 ms). When the filter is tuned to cancel /u/ instead, the SACF reflects /o/.

3.3 Concurrent vowel identification model

In this section the second scheme (time-domain) is developed into a model of concurrent vowel identification, and compared with experimental results. We favor simplicity rather than realism in details of processing or decision process, the aim being to demonstrate that within-channel neural cancellation can be an effective segregation mechanism. Stimuli were processed by a basilar-membrane/hair-cell model [13] with channels evenly spaced along the ERB scale (4 channels per ERB).

An SACF pattern was calculated from which an estimate of the dominant period was derived. This was used to tune an array of cancellation filters, from the output of which a second SACF pattern was derived. The average RMS output/input ratio (residue) of the cancellation filter was also calculated. Based on these elements, two threshold parameters T1 and T2, and a set of template SACFs representing individual vowels, the model predicts the number of vowels reported and their identity. The algorithm is as follows:

- If the residue is less than T1, the model discards the filtered SACF and matches the unfiltered SACF to reference templates. Two possibilities: a) If the ratio of Euclidean distances to the best and second-best templates is greater than T2, the model reports two vowels; b) If the ratio is less than T2, the model reports one vowel: the best match.
- If the residue is greater than T1, the model notes that cancellation was not perfect and reports two vowels. The first is the best match to the unfiltered SACF. The second is the best match to the filtered SACF, *unless* that match produces the same vowel as the first, in which case the model chooses the second-best match to the unfiltered SACF (this rule enforces the constraint that subjects must respond different vowels).

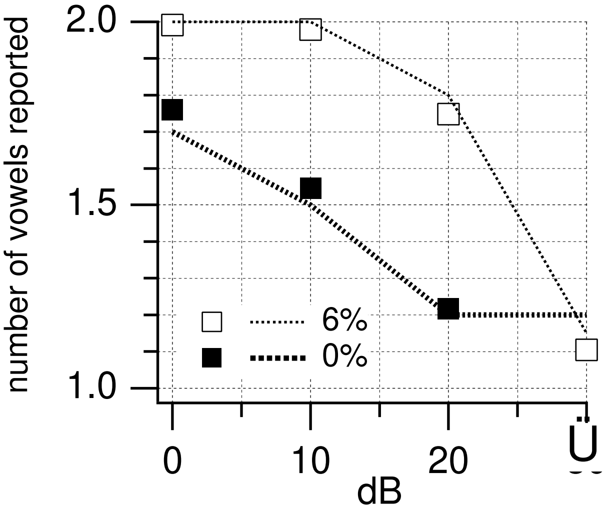

Fig. 8 *Number of vowels reported by subjects (symbols) and predicted by the model (lines) as a function of relative level, at unison (filled symbols, thick line) and* $\Delta F_0 =$ *6% (open symbols, thin lines). Point at right is for single vowels.*

Parameters were adjusted to obtain a good match to the number of vowels reported at unison (T2 = 0.5) and at ΔF_0=6% (T1 = 0.1). Fig. 8 shows the number of vowels reported by subjects (symbols) and predicted by the model (lines). Fig. 9 shows similar data for identification rates. For identification, the match is good at unison but less good at 6%: the model over-estimates performance at low levels (-20 dB) and under-estimates performance at intermediate levels (0, 10 dB). Nevertheless the model correctly predicts a strong ΔF_0 effect at -20 and -10 dB as observed in our experimental data. Overall the match is quite close. One should not make too much of this fact, however, because the model, besides being crude, is deterministic and was tested on only 20 different vowel pairs. The fact that the pattern of results on this small set matched the probabilistic outcome of the experiment is no doubt the result of luck. The important result is that within-channel cancellation can provide effective segregation even when the target is weak.

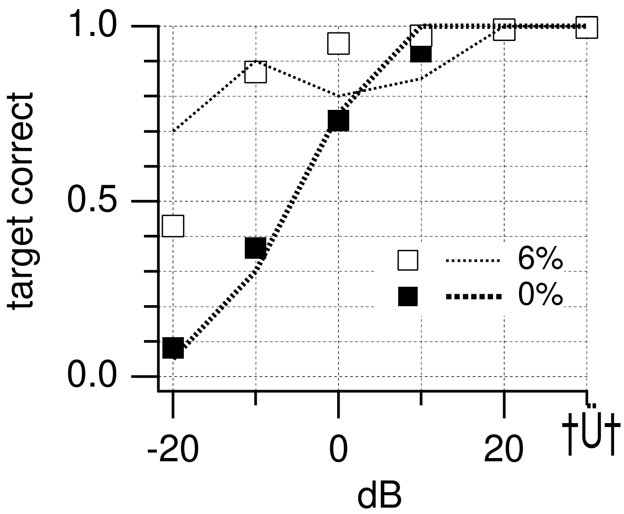

Fig. 9 *Identification rate for subjects (symbols) and predicted by the model (lines) as a function of relative level, at unison (filled symbols, thick line) and* $\Delta F_0 = 6\%$ *(open symbols, thin lines). Point at right is for single vowels.*

4 DISCUSSION

Our experiment reproduced the classic effect of ΔF_0 on identification of mixed vowels, and showed that it persists when the target vowel is weak. A similar conclusion was implicit in the results of Summerfield and Culling [26] who found that a vowel's masking threshold (at 71% correct) fell from 1 dB to -16 dB with an F_0 difference of two semitones. On the other hand McKeown [19] found that ΔF_0 effects tended to vanish at low target levels, possibly because of a floor effect. Our results supported the hypothesis of harmonic cancellation, but offered little evidence that target harmonicity was being used. Nevertheless they did reveal a clear effect of *stimulus* harmonicity, in that double vowels at unison consistently evoked fewer responses than mistuned double vowels (or than inharmonic single vowels [8]).

Within-channel segregation provides an alternative to channel selection as used by Meddis and Hewitt's [22] and other models [10, 4]. Culling and Summerfield's 'Modified Equalization-Cancellation' model, that successfully accounts for many binaural effects, also carries out a form of within-channel cancellation [12]. Within-channel cancellation might be exploited in a variety of fashions, of which the concurrent vowel perception model is an example.

Acknowledgements

The experimental work was carried out at ATR Human information Processing Laboratories, under a research collaboration agreement with the Centre National de la Recherche Scientifique. John Culling provided software for stimulus generation and modelization. Hideki Kawahara, Kiyoaki Aikawa, Minoru Tsuzaki, Stephen McAdams, Jean Laroche and Cecile Marin contributed to the design of the experiment, which Rieko Kubo supervised. Ray Meddis gave useful comments on the model.

References

[1] 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.

[2] Brokx, J. P. L. and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," Journal of Phonetics 10, 23-36.

[3] Bregman, A. S. (1990). Auditory scene analysis. Cambridge, Mass., MIT Press.

[4] Brown, G. J. (1992), "Computational auditory scene analysis: a representational approach.," Sheffield University unpublished doctoral dissertation.

[5] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am., 93, 3271-3290.

[6] de Cheveigné, A. (1994). "Strategies for voice separation based on harmonicity," Proc. ICSLP, Yokohama, 1071-1074.

[7] 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.

[8] de Cheveigné, A., Kawahara, H., Tsuzaki, M. and Aikawa, K. (1996a). "Concurrent vowel segregation I: effects of relative level and F0 difference," in preparation.

[9] de Cheveigné, A., McAdams, S., Marin, M. (1996). "Concurrent vowel segregation II: effects of phase, harmonicity and task," in preparation.

[10] Cooke, M. P. (1991), "Modelling auditory processing and organisation," Sheffield University unpublished doctoral dissertation.

[11] 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.

[12] Culling, J. F. and Summerfield, Q. (1995). "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay." J. Acoust. Soc. Am. 98, 785-797.

[13] Culling, J. F. (1996). "Signal processing software for teaching and research in psycholinguistics." Behavior Research Methods, Instruments, and Computers in press.

[14] Darwin, C. J. (1981). "Perceptual grouping of speech components differing in fundamental frequency and onset-time." QJEP 33A, 185-207.

[15] Graph, J., Hubing N.(1993). "Dynamic time warped comb filter for the enhancement of speech degraded by white gaussian noise,". Proc. IEEE-ICASSP, 339-342.

[16] Green, P.D., Cooke, M.P., and Crawford, M.D. (1995). "Auditory scene analysis and hidden markov model recognition of speech in noise," Proc. IEEE-ICASSP, 401-404.

[17] Hardwick, J., Yoo, C.D., Lim, J.S. (1993). "Speech enhancement using the dual excitation speech model," Proc. IEEE-ICASSP, 367-370.

[18] Lea, A. (1992). "Auditory models of vowel perception," unpublished doctoral dissertation, Nottingham.

[19] McKeown, J. D. (1992). "Perception of concurrent vowels: The effect of varying their relative level," Speech Comm. 11, 1-13.

[20] Meddis, R. (1986). "Simulation of mechanical to neural transduction in the auditory receptor," J. Acoust. Soc. Am. 79, 702-711.

[21] Meddis, R., Hewitt. M.J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification." J. Acoust. Soc. Am. 89, 2866-2882.

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

[23] Palmer, A. R. (1992). "Segregation of the responses to paired vowels in the auditory nerve of the guinea-pig using autocorrelation", in "Audition speech and language", Edited by M.E.H. Schouten, Berlin, Mouton-DeGruyter, 115-124.

[24] Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection." J. Acoust. Soc. Am. 60, 911-918.

[25] Scheffers, M.T.M. (1983), "Sifting vowels," Gröningen University unpublished doctoral dissertation.

[26] Summerfield, Q. and Culling, J.F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency". 124th meeting of the ASA [J. Acoust. Soc. Am. 92, 2317 (A)].

[27] Weintraub, M. (1985), "A theory and computational model of auditory monaural sound separation," Stanford University unpublished doctoral dissertation.

[28] Zissmann, M.A. and Weinstein C.J.(1989). "Speech-state-adaptive simulation of co-channel talker interference suppression.", Proc. IEEE-ICASSP, 361-364.