AUDITORY SEGREGATION OF VOWEL-LIKE SOUNDS WITH STATIC AND DYNAMIC SPECTRAL PROPERTIES

Pierre L. Divenyi[1]
René Carré[2]
and
Alain P. Algazi[1]

In D. P. W. Ellis (Ed.), *IEEE Mohonk Mountain Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 14.1.1-4). New Paltz, N.Y.: IEEE.

ABSTRACT

Experiments were conducted to determine the extent to which a fundamental frequency or formant frequency transition influenced segregation of a simultaneous pair of single-formant harmonic complexes. Results showed that even a minute transition facilitated segregation. The effect was larger for formant frequency than fundamental frequency transitions. It is concluded that dynamic aspects of speech must be taken into account when explaining auditory scene analysis by humans and when designing computational scene analysis methods.

1. INTRODUCTION

One task of auditory scene analysis (Bregman 1991), frequently called upon in everyday life, is performed when the listener has to extract a stream of speech from the cacophony of surrounding noise. Perhaps the most difficult condition for this extraction is when the surrounding noise consists of other streams of speech, i.e., when several individuals speak simultaneously and the auditory system has to achieve perceptual segregation of utterances by different talkers. Such segregation is facilitated by spatial separation between the speakers, although not to the extent that it was initially believed (Divenyi 1995). Fortunately, listeners are capable of following multiple streams of speech even within a single spatial channel – e.g., when conversations in a "cocktail-party" setting are transmitted via a single loudspeaker. Because multiple speech streams only rarely consist of utterances by the same talker, auditory segregation must rely on separating the different voices most probably by extracting the fundamental frequencies f_0 and by grouping together harmonic components that are multiples of the different fundamental frequencies (Bregman, 1991, chapters 3 and 6). The process of grouping has been the focus of an increasing number of studies during the recent years (for a summary, see Darwin and Carlyon 1995). Since fundamental frequency is dominant during vocalic segments of speech, a substantial portion of the studies addressed the question of how simultaneously presented pairs of vowels are identified and segregated (Scheffers 1982; Culling and Darwin 1993; Assmann 1995).

Although much of this research dealt with segregation of steady-state vowels and vowel-like complex sounds, in real speech, vowels change over time with respect to fundamental frequency and formant structure. While the role of f_0 changes in the segregation of simultaneous streams of running speech has been recognized (e.g., McAdams 1989; Gardner, Gaskill et al. 1989; Summerfield and Culling 1992), only scant attention has been devoted to the ways in which a time-varying formant envelope, i.e., vowel quality, will affect segregation [the report by Assmann (1995) is one notable exception]. Interestingly, results of those studies indicate that neither an f_0 variation nor a formant (F_1 and/or F_2) transition will make the vowel in which the change occurred become more identifiable – it is, in fact, the other, steady-state member of the vowel pair that becomes more salient.

2. CRITIQUE OF DOUBLE VOWEL IDENTIFICATION/SEGREGATION EXPERIMENTS

Results of many studies on auditory segregation of vowels must be interpreted with the understanding that the methods most frequently used (i.e., identification of both vowels presented in pairs) have inherent problems. First, the relative perceptual salience of each of two simultaneous vowels will vary from one pair to the next, with the consequence that one or two vowels will dominate the percept of a pair and, consequently, severely bias the ensemble of results. Second, restricting the set of tokens to the ten possible pairs of five English vowels (as customarily done in these studies) makes it possible to observe only the crudest vowel quality effects. [The Dutch vowels used by Scheffers's (1982) study provide a much richer stimulus set.] Third, identification is not synonymous with segregation: A strictly bottom-up model of auditory processing would have to postulate that segregation must precede identification. When, however, one realizes that a salient vowel, e.g., /a/, will likely dominate even a fused percept, the listener may use his/her top-down knowledge to assign a label to this percept without a need for the two vowels to be actually segregated. Fourth, using pairs of vowels with all their formants present makes it impossible to pinpoint the particular vowel feature or features (i.e., formant frequency, formant width, spacing between formants, etc.) that will facilitate segregation. Using dynamic f_0 and/or formant changes is bound to further increase the difficulty of interpreting results.

The experiments reported below extend our inquiry to the general properties of auditory segregation of pairs of simultaneous sounds

and, in particular, the segregation of single-formant vowel-like sounds. Of primary interest to the present experiments is the question of how changes either in the fundamental frequency (with formant peak F_1 held constant) or in the formant peak frequency (with f_0 held constant) will influence segregation in a task in which identification is not required.

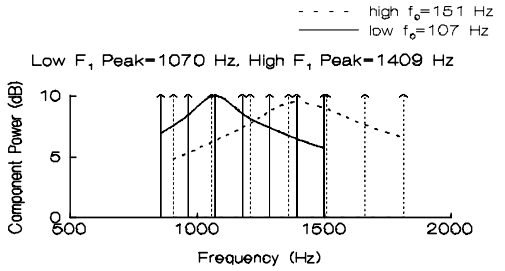

Figure 1: Schematic power spectrum diagram of a typical pair of sinusoidal complexes used in the experiments. Harmonic components of the 107-Hz fundamental f_{0Low} are solid lines, whereas those of the 151-Hz fundamental f_{0High} are broken lines. In this example, the low-f_0 complex also has a lower formant peak F_{1Low} (at 1070 Hz) while the peak of the high-f_0 complex is F_{1Low} (at 1409 Hz). The falloff rate at either side of the peaks is 6 dB/oct.

3. METHODS

2.1 Stimuli

Stimuli for all experiments consisted of 400-ms pairs of simultaneously presented sinusoidal complexes containing 7 (Experiment 1) or 13 (Experiment 2) harmonic components and having different fundamental frequencies. The fundamental frequency of the low-pitch sound f_{0Low} was 107 Hz, whereas that of the higher-pitch sound f_{0High} varied between 120 and 151 Hz. The lowest harmonic component of the low-f_0 sound was always the eighth, whereas that of the high-f_0 one was chosen such as to result in the largest possible total spectral overlap (usually between the fifth and the seventh). The power spectrum of each of the two sounds was shaped: they had their own particular formant peak frequencies F_{1Low} and F_{1High}, above and below which the power of the harmonics was attenuated following a 6-dB/octave falloff. An illustrative example of the spectrum of a complex sound pair is shown in Figure 1. The two sounds of the pair were equalized re/ their overall rms value.

The example in Fig. 1 illustrates stimuli used for the steady-state control conditions. Stimuli for Experiment 1 had the same general spectral structure and constant F_1 peaks, except that the fundamental frequencies of the two sounds included a linear up-down pattern in the temporal center: the f_0 contour of the low-pitch sound was first increased and then decreased, whereas that of the high-pitch sound was first decreased and then increased, as shown in Figure 2a. In contrast, fundamental frequencies in Experiment 2 were constant; the formant peaks, however, had a linear up-down variation in the temporal center of the sounds. The peak frequency of the sound with F_{1High} went down and then up, whereas that of the sound with F_{1Low} first went up and then down, as shown in the diagram in Figure 2b. A single up- or down-transition of frequency had a duration of 50, 100 (as in Fig. 2), or 200 ms, i.e., the complete transition segment had 100, 200, and 400 ms duration for the three conditions, respectively.

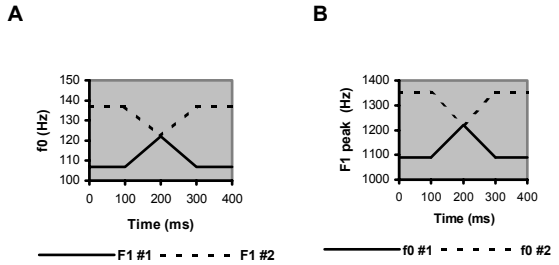

Figure 2: Schematic time diagram of the dynamic frequency changes for sound pair in the two experiments. Panel A: f_0 contours of the two sounds having F_1 peaks F_{1High} and F_{1Low}, randomly assigned to each of the two f_0 contours at each trial. Panel B: F_1 contours of the two sounds having fundamental frequencies f_{0Low} and f_{0High}, randomly assigned to each of the F_1 contours at each trial.

2.2 Procedures

According to our definition (see Divenyi 1995), segregation occurs when, in two simultaneous sounds, the respective stimulus values along at least two dimensions can be correctly associated with one another. In the present experiments, the two dimensions were fundamental frequency (i.e., pitch) and formant peak frequency (i.e., timbral quality). In each experimental condition, stimulus values along one of the dimensions were held constant, while the *difference* between the values of the two simultaneous sounds along the other dimension was varied according to a three-up one-down adaptive two-alternative forced-choice (2AFC) paradigm (Levitt 1971). Thus, in Experiment 1, the F_1 values were held constant at a difference $\Delta F_1/F_1=0.25$ (with the mean F_1 of the two sounds fixed at 1,350 Hz) and the goal was to find a threshold pitch difference $\Delta f_0/f_0$. From trial to trial, the frequency of f_{0High} was varied; the extent of the transitions, however, was always one-half of the frequency difference Δf_0 – i.e., the two frequencies always met at the middle of the stimulus. Similarly, in Experiment 2, the f_0 values were held constant at a difference $\Delta F_1/F_1=0.12$, 0.271, or 0.411 (with the mean F_1 of the two sounds fixed at 113.5, 121.5, or 129 Hz, respectively). From trial to trial, the two frequencies F_1 varied in the opposite direction but the extent of the transitions was, again, always one-half of the frequency difference ΔF_1.

Highly trained subjects (three in Experiment 1 and four in Experiment 2) participated in the study. All of them had

extensive experience in auditory segregation experiments *without* frequency transitions. Some of those data constitute the control conditions for the present study.

4. RESULTS

Weber fractions for fundamental frequency separation just large enough to allow segregation of the two sounds in Experiment 1 are shown in Figure 3. The left-hand panel displays results for the segregation of steady-state pairs of sounds with similar formant frequency separation, i.e., with constant f_0 values. By plotting these threshold values above the "floor" limit, we wish to convey the fact that this particular task was undoable for steady-state pairs of sounds no matter how large an f_0 separation we chose. In contrast, the presence of f_0 transitions made segregation of the two sounds possible, even though (as Fig. 2 clearly shows) the changes *decreased*, rather than increased, the separation of the fundamental frequencies during the transition period. There is also a general, albeit small, trend indicating a more likely segregation for longer transitions.

Results of Experiment 2 are shown in Figure 4 as Weber fractions for formant peak frequency separation just large enough to guarantee segregation, with panels A, B, and C displaying the data for increasing degrees of fundamental frequency separation. Here, too, the left panels illustrate performance for the steady-state formant conditions and indicate that, at all three f_0 separations, segregation was greatly facilitated by the presence of

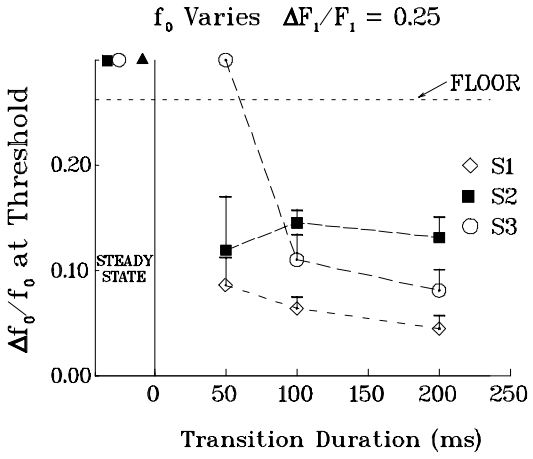

Figure 3: $\Delta f_0/f_0$ Weber fraction thresholds necessary to segregate a pair of harmonic complexes with a fixed $\Delta F_1/F_1$ of 0.25. Results for three subjects. Floor indicates parameters beyond which the task was undoable. Left-hand side displays the results for similar formant frequency and fundamental frequency separations in steady-state sound pairs.

formant transitions. Recall that, as Fig. 2 shows, the separation between formant frequencies was *decreased* during the transition period. Although for some subjects the task is often too difficult (as suggested by the data points beyond the performance floor), the improvement offered by the transitions is indisputable. The transition duration most conducive to segregation is 100 ms at the narrowest f_0 separation (panel A) and ≥100 ms for the others.

5. DISCUSSION

Results of the above experiments clearly show that even a very small linear up-down or down-up frequency transition imposed on either the fundamental frequency or the formant peak frequency of simultaneously presented single-formant harmonic complexes will increase the likelihood of auditory segregation. Comparing the magnitude of the formant frequency and fundamental frequency differences at segregation threshold, it appears that formant frequency transitions are the more important cues for inducing segregation. This means that vowel quality changes can, in fact, lead to segregation as long as the changes take place over a syllabic-length (≥200 ms for the total duration of up-down transitions) interval. These results, when translated into the usual F_1-F_2 plane representation of the first two formant peaks of vowels, can be used to identify the vowel features that listeners may have used, in double-vowel experiments by .McAdams (1989), Summerfield (1992) and others, to arrive at the segregation-plus-identification results these authors reported. Nevertheless, we believe that, with our single-formant harmonic complexes, the segregation we measured was more likely primitive, or "bottom-up", as opposed to the two-vowel experiments cited which, in all likelihood, assessed a combination of primitive (and linguistic experience-derived) schema-driven segregation, using a paradigm that made dissociation of the two processes very difficult.

In conclusion, our results show that dynamic changes in spectrum and/or pitch help listeners segregate speech-like sounds, and suggest that these changes represent important features of speech that, if incorporated into computational scene analysis methods, are likely to reap tangible benefits.

ACKNOWLEDGMENT

This research was supported by the Veterans Affairs Medical Research, by the National Institutes on Aging, and by a NATO Travel Fellowship.

REFERENCES

1. Assmann, P. F. (1995). "The role of formant transitions in the perception of concurrent vowels." *Journal of the Acoustical Society of America 97* : 575–584.
2. Bregman, A. S. (1991). *Auditory scene analysis*. Cambridge, Mass., Bradford Books (MIT Press).
3. Culling, J. E. and J. C. Darwin (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by f0." *Journal of the Acoustical Society of America 93* : 3454-3467.
4. Darwin, C. J. and R. P. Carlyon (1995). Auditory grouping. *Hearing: Handbook of Perception and Cognition, 2nd ed.*, B. C. J. Moore (ed). San Diego, Academic Press: 387-424.

5. Divenyi, P. L. (1995). Auditory segregation of concurrent signals: An operational definition. *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* . New Paltz, New York, Section 2.3, 1-4.

6. Gardner, R. B., S. A. Gaskill, et al. (1989). "Perceptual grouping of formants with static and dynamic differences in fundamental frequency." *Journal of the Acoustical Society of America 85* : 1329-1337.

7. Levitt, H. (1971). "Transformed up-down methods in psychoacoustics." *Journal of the Acoustical Society of America 49* : 467-477.

8. McAdams, S. (1989). "Segregation of concurrent sounds I: Effects of frequency modulation coherence." *Journal of the Acoustical Society of America 86* : 2148-2159.

9. Scheffers, M. T. M. (1982). "The role of pitch in the perceptual separation of simultaneous vowels." *IPO Annual Progress Reports 17* : 41-45.

10. Summerfield, A. Q. and J. F. Culling (1992). "Auditory segregation of competing voices: Absence of FM or AM coherence." *Philosophical Transactions of the Royal Society of London 336* (Series B): 357-366.

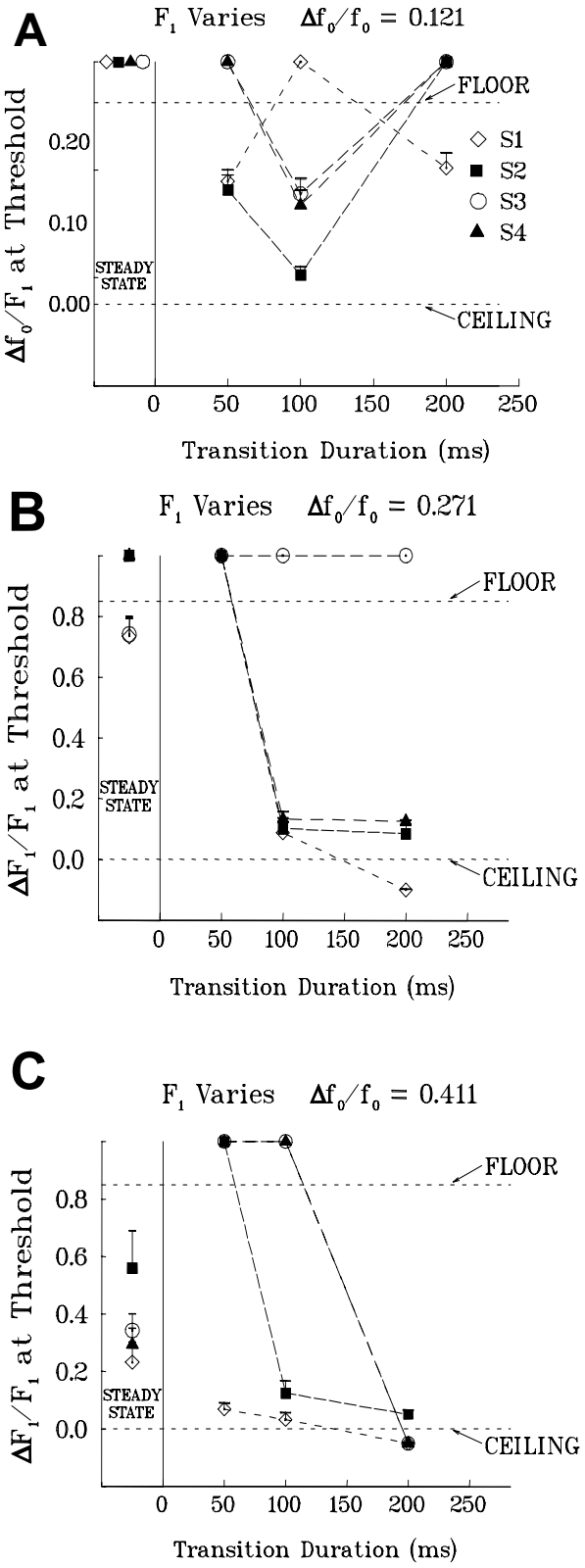

Figure 4: $\Delta F_1/F_1$ Weber fraction thresholds necessary to segregate a pair of harmonic complexes with a given $\Delta f_0/f_0$ – 0.12 (Panel A), 0.271 (Panel B), or 0.411 (Panel C). Results for four subjects. Floor indicates parameters beyond which the task was considered undoable, and ceiling indicates performance better than the minimum threshold measured. Left-hand side displays the results for similar formant frequency-fundamental frequency combinations in steady-state sound pairs.