The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs

Trevor M. Shackleton[a] and Ray Meddis
Speech and Hearing Laboratory, Department of Human Sciences, University of Technology, Loughborough, Leicestershire LE11 3TU, United Kingdom

(Received 25 July 1991; revised 14 November 1991; accepted 20 February 1992)

The ability of subjects to identify both components of a concurrent vowel pair was determined when vowel fundamental frequency difference and an interaural time difference (ITD) in one of the vowels were introduced. In agreement with earlier studies it was found that introduction of a fundamental frequency difference significantly improved detection performance (by 22% for a 1 semitone difference). Identification performance improved as the ITD was made larger but this improvement was small compared to the fundamental frequency difference effect (7% improvement between 0 and 400 μs).

PACS numbers: 43.71.Es, 43.66.Pn, 43.66.Qp

INTRODUCTION

The ability of people to pick out and understand a single conversation in a background babble of many other conversations and noise is remarkable when the poor signal-to-noise ratio is considered. This ability has been described as the "cocktail party effect" (Cherry, 1953). Cherry suggested a number of cues that could be used to perform this task, including voice frequency and spatial separation.

Several studies have shown that there is an important increase in intelligibility when a spatial separation between speech and masker is introduced. Using free-field presentation, a 5- to 9-dB improvement is obtained when speech and masker are separated by 90° (Gelfand *et al.*, 1988; Plomp, 1976). The size of this effect is partly due to head-shadowing, and partly due to binaural processing. A smaller effect (3–7 dB) is found when interaural time differences (ITDs) are used to introduce a difference in lateralization in headphone presentation (Bronkhorst and Plomp, 1988; Carhart *et al.*, 1967,1969; Levitt and Rabiner, 1967).

The role of fundamental frequency difference (Δf_0) in the segregation of speech from background noise has also been extensively studied. Performance improves in the identification of the components of a concurrently presented vowel pair when a Δf_0 is introduced (Assmann and Summerfield, 1989; Chalikia and Bregman, 1989; Culling, 1990; Scheffers, 1983; Zwicker, 1984). A Δf_0 of 1 semitone leads to an improvement of between 15% (Scheffers, 1983; Zwicker, 1984) and 20% (Summerfield and Assmann, 1991) compared to the case with no fundamental frequency difference.

In addition, both Summerfield and Assmann (1991) and Zwicker (1984) have shown that when two concurrent vowels are presented to different ears then there is an improvement of identification performance. However, this improvement could be due either to a reduction in *monaural* interference between the vowels since they pass through different channels, or the improvement could be due to the perceived lateral separation of the vowels.

This paper is intended to achieve two related aims. First, the concurrent vowel identification paradigm is extended to include binaural cues to allow direct comparison with other concurrent vowel results. Second, the question of whether the improvement found by Summerfield and Assmann (1991) and Zwicker (1984) using dichotic vowels was due to binaural effects is answered by introducing a pure binaural cue, namely interaural time differences (ITDs).

I. EXPERIMENT

A. Method and stimuli

The same set of five monophthongal British English vowels (/i/, /ɜ/, /ɑ/, /u/, /ɔ/) as used by Summerfield and Assman (Assmann and Summerfield, 1990; Summerfield and Assmann, 1991) was used. These were generated at a sampling rate of 10 kHz using a Klatt synthesizer (Klatt, 1980). Seven ITDs (± 600, ± 400, ± 200, and 0 μs) and two fundamental frequency differences (Δf_0 of 0 and 1 semitone) were introduced between two concurrently presented vowels.

The vowels were synthesized with the same simulated "vocal effort," so they all had different levels. The mean level was 70 dB SPL (with a range of ± 4.4 dB between vowels). The vowels were 200 ms long with a raised-cosine ramp at onset and offset of 10-ms duration.

One vowel in each pair had a fundamental frequency of 100 Hz and was presented with no ITD. The other vowel in the pair had the same, or a higher frequency and had an imposed ITD. The ITD was created by digitally advancing the vowel waveform to one ear by ITD/2, and delaying the vowel waveform to the other by ITD/2. The entire waveform was advanced or delayed, so there were ITDs in the onsets and offsets of the vowels.

All 25 possible combinations of vowels were used. Each combination of vowels with each ITD and each Δf_0 were repeated once in a session which lasted approximately 30–40

[a] Present address: Laboratory of Experimental Psychology, University of Sussex, Brighton, E. Sussex, BN1 9QG, United Kingdom.

 0001-4966/92/063579-03$00.80

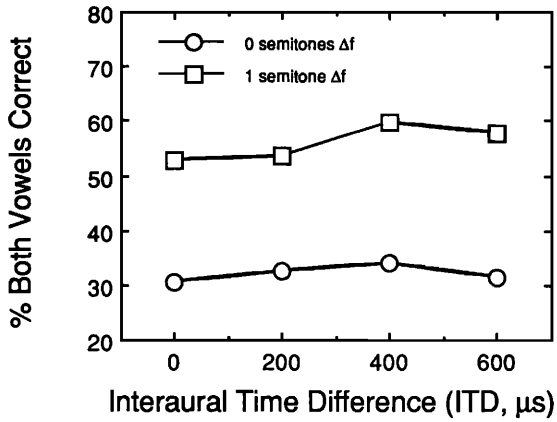

FIG. 1. Variation in concurrent vowel identification performance when an interaural time difference (ITD) is introduced into one of the two vowels. The two curves show performance when both vowels are presented with the same fundamental frequency (circles), and when a fundamental frequency difference of 1 semitone is introduced (squares).

min. Six subjects with normal hearing were used in a single session each. None of the subjects had any experience or training in concurrent vowel identification.

B. Results

The results are shown in Fig. 1. The percentage of trials in which both vowels were correctly identified is shown as a function of ITD with the f_0 difference as a parameter. The results for all subjects are averaged together since although they showed differences in the overall level of performance, they exhibited similar trends. Performance was found to be similar for positive and negative ITDs so these are also shown averaged together. Each data point thus represents the result of 300 trials (excepting 0-μs ITD that represents 150 trials).

The effects of fundamental frequency difference and ITD are significant (ANOVA, Δf_0: $F_{1,5} = 69.55, p < 0.001$; ITD: $F_{3,15} = 3.51, p < 0.05$). The interaction between Δf_0 and ITD is not significant ($F_{3,15} = 0.66$).

The data exhibit a trend for identification performance to increase with increasing ITD from 0 to 400 μs, there is then a small decrease at 600 μs. The increase is slightly larger when Δf_0 is 1 semitone. However, this effect is small (7% between 0 and 400-μs ITD) and is dwarfed by the effect of f_0 difference (22% between 0 and 1 semitone Δf_0).

II. DISCUSSION

There is clearly an improvement in concurrent vowel identification when a binaural cue (namely ITD) is introduced. However, this effect is surprisingly small (7%) when compared with the effect due to fundamental frequency difference (22%). It is unlikely that the small binaural effect is due to an overall ceiling effect for a number of reasons. First, the maximum percentage correct is well below 100%, and some subjects only achieved a maximum of about 30%–40% correct. Second, the improvement in performance for both Δf_0s is similar. It must, therefore, be concluded that the binaural effect is genuinely small.

The overall level of performance is lower than has been reported by others (Summerfield and Assmann, 1991; Zwicker, 1984). This is probably solely due to the use of naive subjects who have not been extensively trained.

How can the small size of this binaural effect be reconciled with the significant binaural effects described in the Introduction? The increase in intelligibility (3–7 dB) due to ITDs in headphone presentation (Bronkhorst and Plomp, 1988; Carhart *et al.*, 1967,1969; Levitt and Rabiner, 1967) is slightly less than the increase in intelligibility (5–9 dB) due to true spatial separation in free-field studies (Gelfand *et al.*, 1988; Plomp, 1976), but both provide a significant increase in intelligibility. Summerfield and Assmann (1991) have also demonstrated a 5-dB improvement in vowel identification threshold when a Δf_0 of 1 semitone was introduced between target vowel and masker vowel. These results are not strictly comparable since different tasks and maskers were used, but there is a discrepancy between the similar performance levels for binaural and pitch cues in the masking studies and the very dissimilar performance in the dual identification study described in this paper.

It is probable that the important identification information in speech sounds is at a higher frequency (above 1 kHz) than the information used to lateralize sounds (below 1500 Hz) when ITDs are imposed upon the fine structure (Carhart *et al.*, 1967, 1969; Zurek, 1992). This means that the binaural and identification information may be weakly coupled together in the task used in this letter, and so a relatively small binaural effect might not be surprising. However, an improvement in recognition of comparable size has been found when different vowels are presented to different ears (Summerfield and Assmann, 1991; Zwicker, 1984), a condition of maximal interaural level difference (ILD). ILDs are binaural cues which are generally regarded to operate in the high-frequency region (> 1500 Hz), so the relatively small size of the effect would appear to be more due to the relative unimportance of binaural cues than to a weak coupling between lateralization information at low frequencies and identification information at high frequencies. In addition, it should be noted that vowel waveforms have a very clear envelope, and that envelope information is useful for lateralization.

The effect may also be small because the ITDs used were ineffective in generating a lateral separation. However, subjects normally reported that they heard the two vowels clearly at different lateralizations when both an ITD and a Δf_0 were introduced. There thus appears to be a clear lateral separation of the vowels due to ITD, but this separation only appears to have a small effect on performance.

These results concur with the observation that binaural fusion of spectral features appears to precede pitch identification (Houtsma and Goldstein, 1972; Raatgever, 1980) or the identification of vowels (Broadbent, 1955; Broadbent and Ladefoged, 1957; Cutting, 1976). This indicates that spectral features seem to be more readily grouped by monaural cues such as onset time and common fundamental than by ear of origin.

Whilst the difference in size between the binaural and pitch effects appears substantial and is thus of theoretical

importance, it is possible that the ecological relevance of the finding is less significant (P. M. Zurek, personal communication). The incremental improvement in vowel identification for separations greater than one semitone is very small compared with the improvement between zero and one semitone (Assmann and Summerfield, 1990; Summerfield and Assmann, 1991; Scheffers, 1983). Voice pitches are very rarely within one semitone of each other, so identification nearly always operates with the maximal pitch cue in operation. Thus modulations in voice pitch are unlikely to lead to further improvement. However, the ITD effect appears to increase fairly uniformly throughout the range studied, so we could expect an improvement due to binaural cues to occur whenever speakers move further apart. In other words, although the experiments reported in this paper suggest that binaural cues are less significant than pitch cues it is possible that in natural situations binaural cues may still allow improvements in performance as speakers move apart whereas the pitch cue is similar for different speakers.

III. CONCLUSION

The introduction of an ITD difference between two vowels presented concurrently leads to an improvement in the identification of both vowels of up to 7%. The ITD effect is small compared with the effect due to a fundamental frequency difference (22%). This leads us to conclude that pitch cues are more salient than binaural cues, at least for this task.

ACKNOWLEDGMENTS

We would like to thank Quentin Summerfield, Chris Darwin, John Culling, and Andrew Lea for a number of helpful comments during the course of this study. We would also like to thank P. M. Zurek, an anonymous reviewer, Brian Moore, and Michael Hewitt for useful comments about an early version of this paper. This work was supported by the Science and Engineering Research Council (UK) Image Interpretation Initiative (Grant No. GR/E 88240).

Assmann, P. F., and Summerfield, A. Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–337.

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.

Broadbent, D. E. (**1955**). "A note on binaural fusion," Q. J. Exptl. Psychol. **7**, 46–47.

Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708–710.

Bronkhorst, A. W., and Plomp, R. (**1988**). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am. **83**, 1508–1516.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Release from multiple maskers: Effects of interaural time disparities," J. Acoust. Soc. Am. **45**, 411–418.

Carhart, R., Tillman, T. W., and Johnson, K. R. (**1967**). "Release of masking for speech through interaural time delay," J. Acoust. Soc. Am. **42**, 124–138.

Chalikia, M. H., and Bregman, A. S. (**1989**). "The Perceptual Segregation of Simultaneous Auditory Signals: Pulse Train Segregation and Vowel Segregation," Percept. Psychophys. **46**, 487–496.

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

Culling, J. (**1990**). "Perception of simultaneous vowels: effects of across-formant inconsistencies in f_0," Brit. J. Audiol. **24**, 194.

Cutting, J. E. (**1976**). "Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening," Psychol. Rev. **83**, 114–140.

Gelfand, S. A., Ross, L., and Miller, S. (**1988**). "Sentence reception in noise from one versus two sources: Effects of aging and hearing loss," J. Acoust. Soc. Am. **83**, 248–256.

Houtsma, A. J. M., and Goldstein, J. L. (**1972**). "The central origin of the pitch of complex tones: Evidence from musical interval recognition," J. Acoust. Soc. Am. **51**, 520–529.

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

Levitt, H., and Rabiner, L. R. (**1967**). "Binaural release from masking for speech and gain in intelligibility," J. Acoust. Soc. Am. **42**, 601–608.

Plomp, R. (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)," Acustica **34**, 200–211.

Raatgever, J. (**1980**). "On the binaural processing of stimuli with different interaural phase relations," Doctoral dissertation, Technische Hogeschool Delft, The Netherlands.

Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Doctoral dissertation, Groningen.

Summerfield, Q., and Assmann, P. F. (**1991**). "Perception of concurrent vowels: Effects of pitch-period asynchrony and harmonic misalignment," J. Acoust. Soc. Am. **89**, 1364–1377.

Zurek, P. M. (**1992**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston, MA) (in press).

Zwicker, U. T. (**1984**). "Auditory recognition of diotic and dichotic vowel pairs," Speech Commun. **3**, 265–277.