Limits to the role of a common fundamental frequency in the fusion of two sounds with different spatial cues

C. J. Darwin[a] and R. W. Hukin
Department of Psychology, University of Sussex, Brighton, BN1 9QG, United Kingdom

(Received 17 November 2003; revised 16 April 2004; accepted 18 April 2004)

Two experiments establish constraints on the ability of a common fundamental frequency (F0) to perceptually fuse low-pass filtered and complementary high-pass filtered speech presented to different ears. In experiment 1 the filter cut-off is set at 1 kHz. When the filters are sharp, giving little overlap in frequency between the two sounds, listeners report hearing two sounds even when the sounds at the two ears are on the same F0. Shallower filters give more fusion. In experiment 2, the filters' cut-off frequency is varied together with their slope. Fusion becomes more frequent when the signals at the two ears share low-frequency components. This constraint mirrors the natural filtering by head-shadow of sound sources presented to one side. The mechanisms underlying perceptual fusion may thus be similar to those underlying auditory localization. © *2004 Acoustical Society of America.* [DOI: 10.1121/1.1760794]

PACS numbers: 43.66.Pn, 43.66.Rq [PFA] Pages: 502–506

I. INTRODUCTION

In a well-known paper, Broadbent and Ladefoged (1957) demonstrated the importance of a common fundamental frequency (F0) in the perceptual fusion of sounds with different spectral composition presented to different ears. They played the first formant of a synthetic sentence to one ear of their listeners, the second formant to the other ear, and asked how many voices listeners heard and where they were located. When the two formants were excited by pulses at the same F0, the majority of listeners reported hearing a single voice (13/18) in a single place (15/18), but when the two formants were excited by pulses with different F0s (125 vs 135 Hz), the majority of listeners heard two voices (15/18) in two places (12/18). The ability of a common F0 to fuse sounds with different spectral content across the two ears had previously been noted by Fletcher, following a suggestion by Arnold (Fletcher, 1953 p 216) and by Broadbent (1955). Fletcher describes the fusion that occurred when speech that had been high-pass filtered at 1 kHz was presented to one ear, with the complementary low-pass filtered speech to the other ear (Fletcher also observed that a similar fusion did not occur with polyphonic music). Broadbent produced a more extreme manipulation of speech, with low-pass filtered speech at 450 Hz (-18 dB/oct) to one ear and the same speech high-pass filtered at 2000 Hz to the other ear. Of 18 listeners, 14 reported hearing one voice rather than two. Broadbent comments that the common spectral envelope across the two ears might be responsible for the perceived fusion.

These early observations provided the starting point for a number of papers investigating the effect on the intelligibility of speech of varying the fundamental frequency relations within and between speech sounds (Cutting, 1976; Darwin, 1981; Scheffers, 1983; Assmann and Summerfield, 1989; Assmann and Summerfield, 1990; Culling and Darwin, 1993; Culling and Darwin, 1994; Bird and Darwin, 1998). The Broadbent and Ladefoged original observation on the number of sound sources that listeners hear has received less attention, although it has been confirmed with syllabic sounds where the output of the first-formant filter was led to one ear and that of the second-formant to the other ear (Darwin, 1981).

In setting up demonstrations of the fusion across the ears of bands of speech on a common F0, we had noticed that when there was no spectral overlap between the sounds played to the two ears, fusion was less likely to occur than when there was greater spectral overlap. This observation is interesting, not only because it suggests that there might be limits to the fusion by F0 reported by Broadbent and Ladefoged, but also because it might provide a link between observations on auditory fusion and the extensive literature on auditory localization.

The following experiments explore how fusion depends on spectral overlap between the sounds presented to each ear and demonstrate that sounds are more likely to fuse when they share low-frequency components. This constraint mirrors the diffraction of low-frequency (but not high-frequency) sound around the head.

II. EXPERIMENT 1

The Broadbent and Ladefoged speech sounds were prepared using Walter Lawrence's PAT (Parametric Artifical Talker) synthesizer (Lawrence, 1953). PAT consisted of analogue resonator circuits that filtered a periodic electrical laryngeal signal. Each ear in the Broadbent and Ladefoged experiment thus received the output of a simple resonant filter. Figure 1 shows the transfer function of two such resonators (following Fant, 1970 p. 54 Eq. 1.3-7), one tuned to 800 Hz with a bandwidth of 90 Hz and the other tuned to 1400 Hz with a bandwidth of 150 Hz. Below the axis is shown the negative of the absolute difference between the two functions. It is clear that there is considerable spectral

[a]Correspondence and proofs to C. J. Darwin, Department of Psychology, University of Sussex, Brighton BN1 9QG, England. Electronic mail: cjd@biols.susx.ac.uk

 0001-4966/2004/116(1)/502/5/$20.00

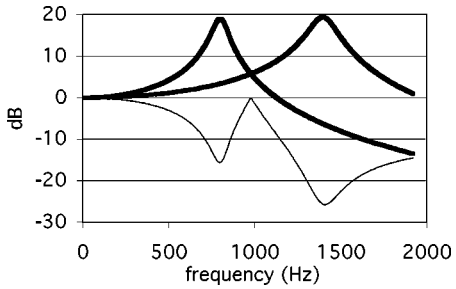

FIG. 1. Transfer functions for two single-formant resonators at 800 Hz and 1400 Hz with bandwidths of 90 and 150 Hz, respectively. The thin line shows the difference in level between the two functions.

overlap between the two. Although the stimuli used by Broadbent (1955) had considerably less spectral overlap than those used by Fletcher or by Broadbent and Ladefoged (see above), the frequency region around 950 Hz would have had the same level in both ears (with an attenuation of about 19 dB) and the high-pass band would have had relatively little energy because of its high lower-frequency limit.

In Experiment 1 we ask how listeners' judgements of the number of sound sources change when the speech to each ear is filtered through complementary high- and low-pass filters whose 6-dB cut-off frequency is fixed at 1 kHz and whose steepness is systematically varied. We also included an additional set of conditions where the high- and low-frequency signals were each sent to both ears, but with complementary interaural time differences (ITDs) in order to investigate whether the basic phenomenon reported by Broadbent and Ladefoged is also obtained using ITDs rather than their dichotic infinite interaural level difference. A number of studies have recently demonstrated the weaker effect of simultaneous auditory grouping by ITD than by infinite ILD (Culling and Summerfield, 1995; Darwin and Hukin, 1997; Drennan *et al.*, 2003).

A. Stimuli and procedure

Two sentences from a speech corpus (Bolia *et al.*, 2000) were used, one spoken by a woman (Talker 4: ''Ready Charlie, go to blue one now''), and one by a man (Talker 5: ''Ready Ringo, go to red six now''). The sentences were first low-pass filtered at 8 kHz. In the simplest condition (dichotic, Same F0), either the male or female sentence was resynthesized with no change to its F0 using the Praat 3.9 (Boersma and Weenink, 1996) implementation of PSOLA (Moulines and Charpentier, 1990). Then a low-pass version was played to one ear of a listener at the same time as a high-pass version was played to the other ear. The filtering was carried out in the frequency domain using Praat's implementation of a Hann filter, which produced (symmetrically around the cut-off frequency) a linear attenuation of the sound on linear frequency and amplitude scales. The cut-off frequency, in this case 1 kHz, is defined as the 6-dB (50% linear) attenuation point of the filter, and is the frequency at which the complementary high- and low-pass filters cross. The total width of the linear attenuation zone varied in 400-Hz steps from 200 Hz to 1800 Hz.

The sounds of two further conditions had different F0s in their low-pass and high-pass parts. The Low–High condition had the lower F0 in its low-pass part and the higher F0 in its high-pass part. The High–Low condition was the opposite. The changes to F0 were made on the intact original sentences again using Praat's PSOLA implementation. For the lower F0 sounds, the F0 was lowered by 4% from its original value, and for the higher F0 it was raised by 4%, giving an overall F0 difference of a little over 8% of the lower value.

In the ''Dichotic'' set of conditions, the low-pass and high-pass parts of a sentence were played to different ears (low-pass always to the left ear). In the ''ITD'' set of conditions, both parts were played to both ears but with an ITD of ± 571 μs applied so that the low-pass part led on the left ear and the high-pass led on the right.

Eight audiometrically normal listeners who had the general experience of taking psychoacoustic experiments, though not of this type, listened to 10 replications of each stimulus in a pseudo-random order in an audiometric booth over Sennheiser 414 headphones. They were asked to indicate on each trial whether they heard one (fused) voice or two. The presentation gain produced a level for the low-pass sound (1000-Hz filter transition width) of 62-dB SPL.

B. Results

The results for the dichotic and ITD presentations are shown in the upper and lower panels of Fig. 2, respectively, as the percentage of trials on which listeners heard a single, fused sound. For both dichotic and ITD presentation, sounds that had a different F0 in their low- and high-pass parts (triangles and squares) were heard as fused on less than 25% of occasions, with the female voice (filled symbols) being heard as less fused than the male (unfilled). There were no systematic changes with the filter transition width.

However, sounds that had the same F0 in both parts (circles) showed a different response pattern. With dichotic presentation (as in the original Broadbent studies) sounds that had been filtered through filters with wide transitions (>500 Hz), were heard as fused, whereas those from filters with narrower transitions were not heard as fused. A repeated-measures ANOVA on the dichotic data (with the scores of the two sub-classes of different F0 averaged) showed a highly significant interaction of ''same vs different F0'' with ''filter transition width'' ($F_{4,28}=44.3, p<0.0001$) which itself interacted only weakly with talker gender ($F_{4,28}=4.0, p<0.05$). These results replicate the Broadbent and Ladefoged result described above, but only for wide filter transitions. For narrow filter transitions, a common F0 is insufficient to give the impression of a single sound source. With ITD presentation, all the sounds with the same F0 were heard as fused more than 75% or so of the time. A repeated-measures ANOVA on the ITD data (with the scores of the two sub-classes of different F0 averaged) showed a highly significant effect of same vs different F0 ($F_{1,7}=53.3, p<0.0002$), but no other main effects or interactions.

C. Discussion

This experiment has confirmed one aspect of the Broadbent and Ladefoged results, namely that when two different

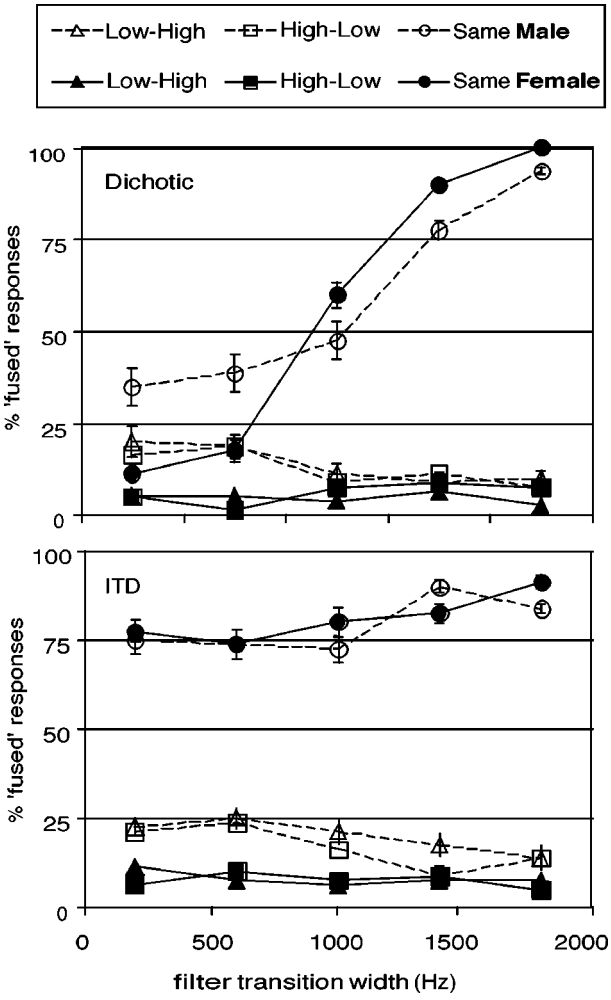

FIG. 2. Percentage of single voice (fused) responses (± 1 s.e.m) in Experiment 1. The upper panel shows data for dichotic presentation of a 1-kHz high-pass and low-pass filtered version of a sentence from a male (open symbols) or female (closed symbols) voice as a function of the width of the linear (in amplitude and frequency) skirts of the filters which crossed at -6 dB. The ''Low–High'' condition had the lower F0 in the low-pass part and the higher F0 in the high-pass part. The ''High–Low'' condition had the opposite assignment. Some error bars fall within their symbols. The lower panel shows data from similar sounds presented with ITDs of ± 571 μs rather than dichotically.

frequency bands are led to opposite ears, fusion is more likely when the sounds are on the same F0 than when they are on different F0s. But the experiment has also qualified this conclusion: such fusion only occurs for frequency bands that have been obtained by passing the original speech through relatively shallow filters. With steeper filters, listeners consistently report two sound sources even when the bands share a common F0.

The need for shallow filters may be because some frequency components must be shared between the two ears for fusion to occur, or it may be due to the need to share specific frequencies (such as the low-frequency region that is dominant for auditory localization). This question is addressed in the second experiment.

The first experiment also examined the fusion of different frequency bands that were played both to each ear but with different interaural time differences. Previous work on auditory grouping has indicated that ITDs provide at best only a weak basis for auditory grouping compared with grouping by ear of presentation (Culling and Summerfield, 1995; Darwin and Hukin, 1997). In the present results we extend this conclusion to judgements of auditory fusion: sounds on the same F0 presented with different ITDs were judged as fused regardless of the width of the filters through which they had been passed. This experiment thus shows that grouping by common F0 overrides potential separation by an ITD of over 500 μs. This result complements previous findings that a difference in F0 is more salient than a difference in ITD at improving the identification of simultaneous pairs of vowels (Shackleton and Meddis, 1992).

III. EXPERIMENT 2

In the second experiment, we vary the cut-off frequency as well as the transition-width of the filters used to generate the low- and high-pass versions of the sentences. The reason for varying both these parameters is to distinguish an explanation in terms simply of filter sharpness from one that requires frequency overlap between the ears in a particular frequency region such as, for example, the dominant region for localization (Raatgever, 1980; Wightman and Kistler, 1992).

A. Stimuli and procedure

The stimuli and procedure were similar to those used in the first experiment, except that there were 5 different cut-off frequencies of the low-/high-pass filter (600, 800, 1200, 1400, 2000 Hz), and presentation was only dichotic. Each cut-off frequency of the filter had the same 5 transition-widths used in the previous experiment. The transition-widths were thus constant on a linear scale, and did not increase in proportion to the filter cut-off frequency.

B. Results

This experiment replicates the dichotic results from the first experiment. For the 800-Hz and 1200-Hz cut-off frequencies (which are the most similar to the 1000-Hz cut-off of the first experiment) there are very few fusion responses when the two pass-bands have different F0s, but when they have the same F0, fusion responses increase as the filter transition width increases.

More generally, as in the first experiment, listeners reported very little fusion for sounds that had a different F0 in the low-pass and high-pass parts: only the male sentence with the highest (2 kHz) cut-off frequency approached 30% fusion responses.

By contrast, the sounds that had the same F0 in both the low- and high-pass parts showed high levels of fusion in some conditions. The percentage of fusion responses for sounds on the same F0 are shown separately for the male and female sentences in Fig. 3.

The data (with the scores of the two sub-classes of different F0 averaged) were subjected to a repeated measures ANOVA which gave a substantial three-way interaction between ''cut-off frequency,'' ''same vs different F0'' and ''filter transition width'' ($F_{16,112}=9.2, p<0.0001$) which weakly interacted with talker gender ($F_{16,112}=3.0, p<0.05$).

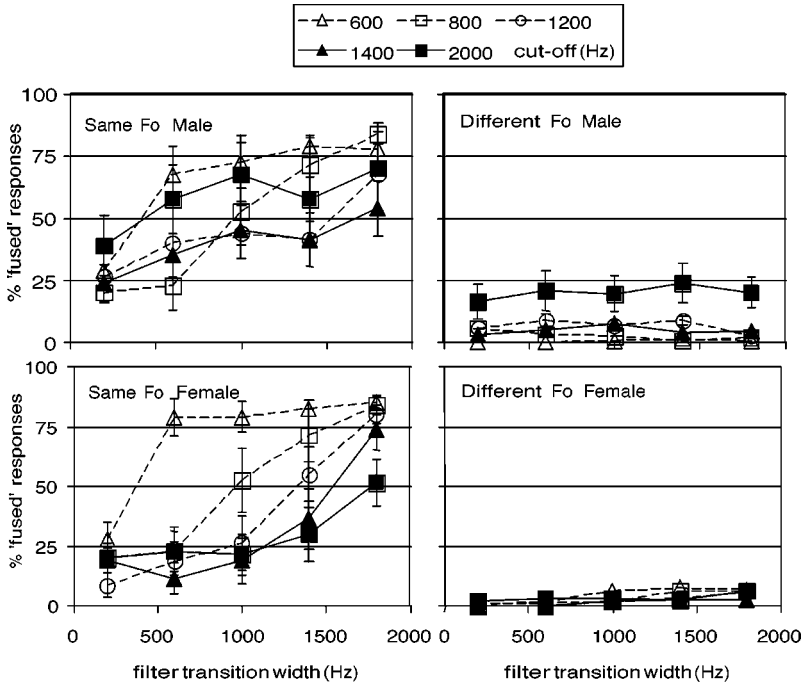

FIG. 3. Percentage of single voice (fused) responses (± 1 s.e.m) in Experiment 2 for sounds on the same F0 as a function of the filter transition width for 5 different filter cut-off frequencies. The upper panels show data from the male talker, the lower from the female; the left column shows data from conditions on the same F0, the right from those on a different F0.

The data from the female talker when the F0s were the same are more orderly than the male and show two trends. First, as in the previous experiment, fusion generally increases with increasing filter transition width. Second, fusion increases as the cut-off frequency of the filter is decreased. So, for example, for a transition width of 500 Hz, fusion responses are less than 25% for filter cut-offs of 1200 to 2000 Hz, but increase to over 75% with a cut-off of 600 Hz. Viewed another way, the higher the cross-over frequency of the filter, the wider must be the filter transition to give fusion.

The male same-F0 data show similar trends to the female, with the exception that the highest filter cross-over frequency 2000 Hz gives substantially more fusion responses than do the female data. The reason for this is not clear, but may reflect weaker pitch information from the high-numbered harmonics of the low-pitched male voice in the region above 2 kHz compared with the lower-numbered harmonics in the same frequency region for the female voice (Houtsma and Smurzynski, 1990). The different phonetic content of the two voices may also have been a factor, giving different distributions of energy between the two pass bands, and this variable needs to be explicitly controlled in future systematic comparisons of different pitched or different gender voices.

C. Discussion

The main result of this experiment is that a broader filter transition region is required for fusion as the cross-over frequency between the low-pass and high-pass sounds is increased. The implication of this result is that successful fusion requires that the high-pass stimulus contain sufficient low-frequency energy. If we consider sounds at around the 50% threshold for fusion responses in the female data in Fig. 3, then the high-pass component of these threshold sounds generally starts to show some energy above about 400 to 600 Hz, and would therefore have substantial energy at slightly higher frequencies—in the dominance region for localization (Raatgever, 1980; Wightman and Kistler, 1992). The level difference between the two ears as a function of frequency is shown in Fig. 4 for filters at these 50% threshold frequencies (for the female voice). With the exception of the highest filter cut-off, all the filters at threshold show overlap of frequencies in the frequency region around 600–700 Hz.

IV. GENERAL DISCUSSION

These two experiments have shown that although a common F0 may be a necessary condition to ensure binaural fusion of two different frequency bands led to opposite ears, it is not a sufficient condition. If the frequencies of a sentence below 1 kHz are played to one ear, and those above 1 kHz to the other, listeners will either report hearing one or two sound sources depending on whether the cross-over filter

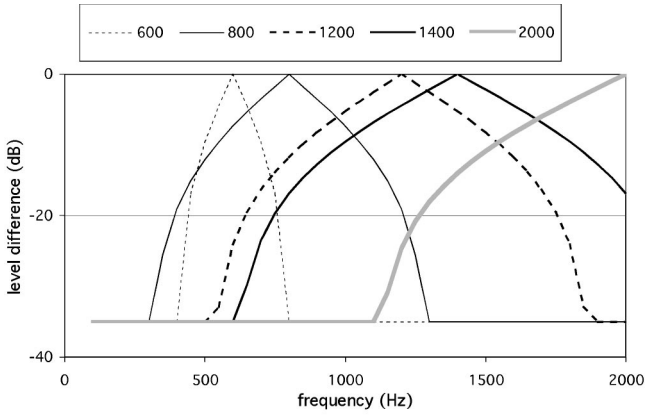

FIG. 4. Level differences between low-pass and high-pass filter transfer functions for female-voice stimulus conditions in Experiment 2 that gave 50% fused responses. The filter widths that corresponded to 50% fused responses were estimated from the average data across listeners at each cut-off frequency, and transfer functions for those filter widths calculated that were linear in frequency and amplitude.

has shallow or steep skirts, respectively. As the cross-over frequency is moved from lower to higher frequencies, shallower skirts to the filters are needed to produce fusion.

A possible explanation for these effects is that for fusion to occur, the high-pass sound must have sufficient energy in the dominance region for lateralization. This constraint may reflect the natural constraint on real sound sources that the head produces a darker acoustic shadow for high frequencies than for low; consequently although it is natural to encounter sounds at one ear from which the high frequencies have been removed (by the head shadow), it is not natural to encounter sounds at one ear from which the low frequencies have been removed. The mechanisms of binaural fusion may be sensitive to this constraint and produce the percept of a separate sound source for a sound that, although likely to be from the same sound source as a low-frequency sound at the other ear by virtue of their sharing a common F0, has too little low-frequency energy. The unity of the resulting percept would then be a trade-off between these two opposing factors.

ACKNOWLEDGMENTS

The research was supported by a grant from the UK Medical Research Council to the first author. This data was initially presented at the Short Papers Meeting of the British Society for Audiology, 2001.

Assmann, P. F., and Summerfield, A. Q. (**1989**). ''Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency,'' J. Acoust. Soc. Am. **85**, 327–338.

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

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

Boersma, P., and Weenink, D. (**1996**). ''Praat, a system for doing phonetics by computer, version 3.4,'' Institute of Phonetic Sciences, University of Amsterdam, Vol. 132, pp. 1–182, www.praat.org

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). ''A speech corpus for multitalker communications research,'' J. Acoust. Soc. Am. **107**, 1065–1066.

Broadbent, D. E. (**1955**). ''A note on binaural fusion,'' Q. J. Exp. 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.

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

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

Culling, J. F., and Summerfield, Q. (**1995**). ''Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay,'' J. Acoust. Soc. Am. **98**, 785–797.

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

Darwin, C. J. (**1981**). ''Perceptual grouping of speech components differing in fundamental frequency and onset-time,'' Q. J. Exp. Psychol. **33A**, 185–208.

Darwin, C. J., and Hukin, R. W. (**1997**). ''Perceptual segregation of a harmonic from a vowel by inter-aural time difference and frequency proximity,'' J. Acoust. Soc. Am. **102**, 2316–2324.

Drennan, W. R., Gatehouse, S., and Lever, C. (**2003**). ''Perceptual segregation of competing speech sounds: the role of spatial location,'' J. Acoust. Soc. Am. **114**, 2178–89.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (The Hague, Mouton).

Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, New York).

Houtsma, A. J. M., and Smurzynski, J. (**1990**). ''Pitch identification and discrimination for complex tones with many harmonics,'' J. Acoust. Soc. Am. **87**, 304–310.

Lawrence, W. (**1953**). ''The synthesis of speech from signals which have a low information rate,'' in *Communication Theory*, edited by W. Jackson (Butterworths Scientific, London, England).

Moulines, E., and Charpentier, F. (**1990**). ''Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones,'' Speech Commun. **9**, 453–467.

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

Scheffers, M. T. (**1983**). ''Sifting vowels: Auditory pitch analysis and sound segregation,'' Ph.D. dissertation, Groningen University, The Netherlands.

Shackleton, T. M., and Meddis, R. (**1992**). ''The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs,'' J. Acoust. Soc. Am. **91**, 3579–3581.

Wightman, F. L., and Kistler, D. J. (**1992**). ''The dominant role of low-frequency interaural time differences in sound localization,'' J. Acoust. Soc. Am. **91**, 1648–1661.