

Vocal Tract Size Normalization in the Perception of Stop Consonants*

Timothy C. Rand⁺
Haskins Laboratories, New Haven

One type of acoustic variation confronting a listener stems from the fact that vocal tract dimensions vary from speaker to speaker. Speakers with smaller vocal tracts generally produce speech with higher formant frequencies. This is apparent in Peterson and Barney's (1952) formant frequency data for vowels produced by men, women, and children.

Ladefoged and Broadbent (1957) have demonstrated context effects on vowel perception, where certain of the variations in the contextual material reflected variations in vocal tract size. The listeners interpreted vowels and preceding carrier phrases as if they had been produced by the same sized vocal tract. These results establish the existence of vocal tract normalizing functions in the human speech perception mechanism. Additional support is provided by Darwin's (1971) demonstration of a right-ear effect when vowels are presented dichotically and vocal tract size is varied from trial to trial.

Consonant perception differs markedly from vowel perception, particularly in the case of the stop consonants. A stop displays a significant lack of acoustic invariance when it occurs together with different vowels. For the place of articulation dimension, this context-conditioned variation has been explained with reference to acoustic loci to which the second formant transition "points" (Delattre et al., 1955). The loci, which are not directly realized in the acoustic speech signal, can be considered to correspond to the resonant frequencies of the occluded vocal tract. Stevens and House (1956) used a vocal tract analog synthesizer to test this hypothesis by measuring resonances when the tract was appropriately constricted. Their findings for second formant loci agree well with the experimental results of the Haskins group.

If the locus concept is to be related to the resonance of the occluded vocal tract, then it is to be expected that this resonant frequency will vary with vocal tract size. Fourcin (1968) demonstrated that prior context influences the perception of synthetic whispered consonant-vowel syllables. Fourcin interpreted his results as a shift in consonantal locus depending upon whether a precursive "hallo" was spoken by a man or a child.

* Paper presented at the 81st meeting of the Acoustical Society of America, Washington, D.C., 20-23 April 1971.

⁺Also the University of Connecticut, Storrs.

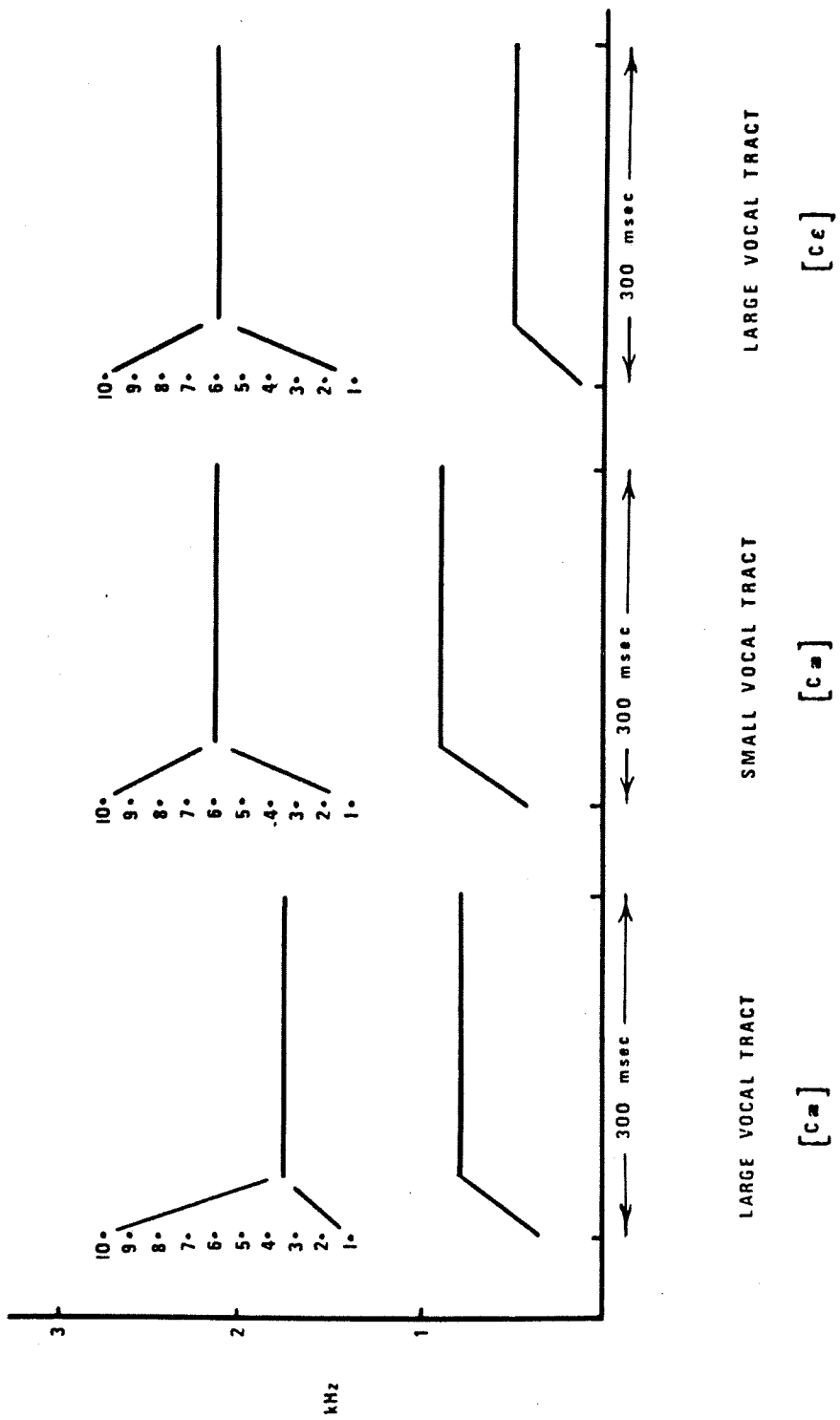


Fig. 1

A further test of locus shift would be to see whether the second formant transitions of CV syllables can cue different consonants depending upon vocal tract information carried on the syllable itself.

The stimuli used in the present experiment (Fig. 1) were thirty two-formant synthetic CV syllables. All syllables were of 300-msec duration and were produced on the Haskins parallel formant synthesizer. The stimuli fall into three groups, based upon the characteristics of the vocalic portion of the syllable. In the first group, the steady-state vowel formants are appropriate to [æ] as rendered by an adult speaker. The vowel formants of the second group of stimuli are related to those of the first group by a multiplicative constant (approximately 1.2); these are roughly appropriate to a child's rendition of [æ]. The vowel of the third stimulus group is roughly [ɛ] as rendered by an adult. Groups 2 and 3 are related in that they have identical second formants.

All stimuli have initial rising first formant transitions to serve as a cue for voiced manner of production. As regards second formant transitions, ten equally spaced frequencies were chosen, ranging from 1386 to 2762 Hz, to serve as initial transition frequencies.

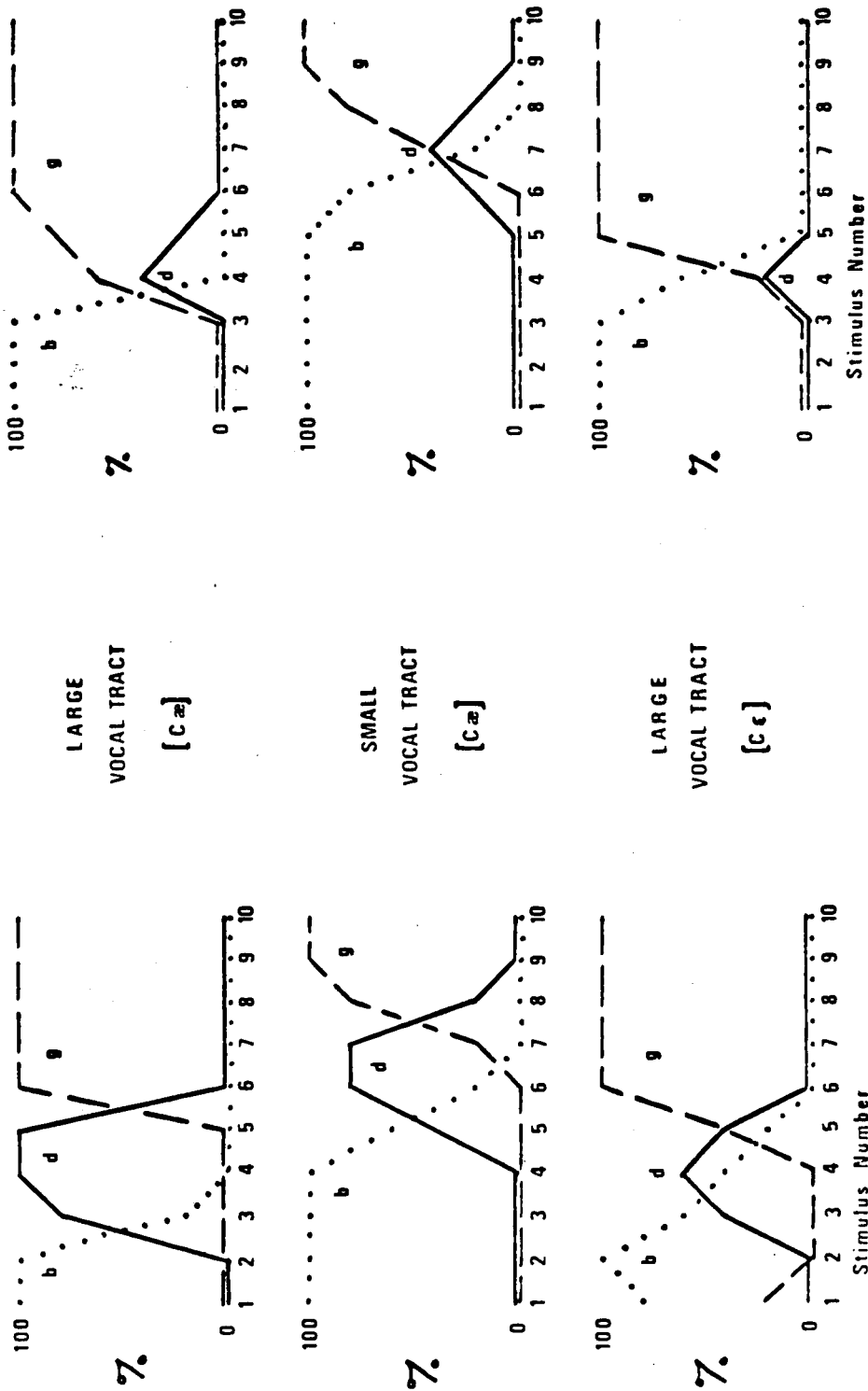
A tape was prepared with five repetitions of each stimulus (150 items total) in a random sequence. Listeners were told they would hear synthetic CV syllables and were instructed to write down the initial consonant for each.

Figure 2 displays identification functions for two subjects. The functions are plotted separately for the three classes of stimuli. In each case, the ordinate is percent identification and the abscissa is the stimulus number, indicating the starting point of the second formant transition. There is a clear tendency for the consonant categories to occur at higher frequencies for the small vocal tract stimuli than for both varieties of large vocal tract stimuli.

Another way to observe how vocal tract size affects place identification is to compute the mean [d] response on the ten-point scale corresponding to initial second formant transition frequency. For example, for the subject whose data are displayed on the left in Figure 2, the mean [d] responses are 4.1, 6.4, and 4.0 for the three stimulus groups going from top to bottom. These means were computed for ten subjects and are displayed in Figure 3.

In Figure 3, the "S's" indicate the mean [d] response for each subject for the Small vocal tract stimuli; similarly, the "L's" indicate the mean values for the two conditions involving Large vocal tract stimuli. The important feature of these results is that, for all ten subjects, small vocal tract stimuli produced the response [d] for higher values of second formant transition frequency than did the large vocal tract stimuli.

Speech utterances are effective communicators whether they originate from small children or large adults. Normalization for vocal tract size differences during the process of speech perception amounts to a cancellation of one type of acoustic variance that is introduced during the production phase of the speech chain. One way to characterize this ability on the part of a listener is to say that he somehow extracts or reconstructs the speaker's



IDENTIFICATION

Fig. 2

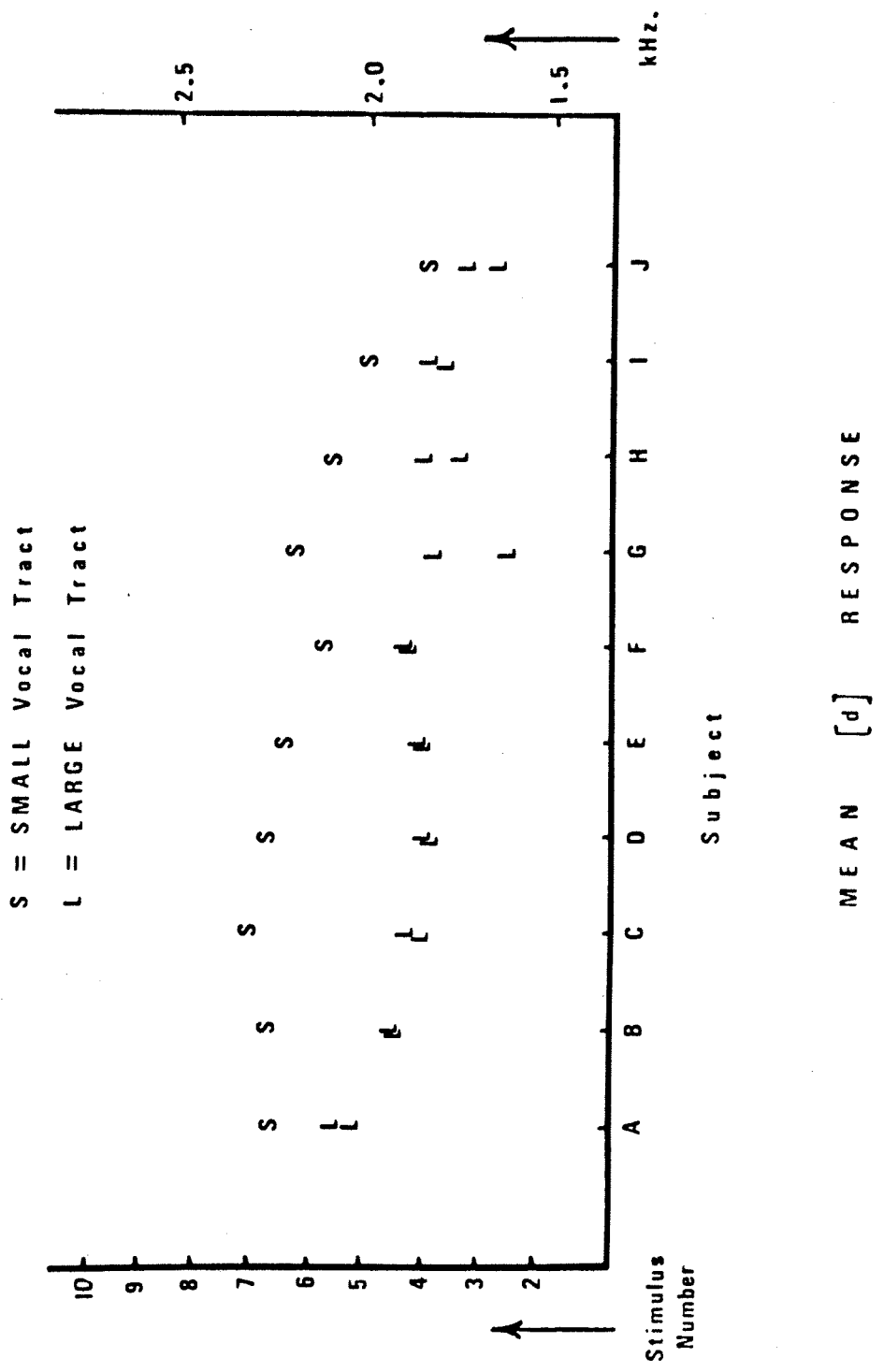


Fig. 3

articulatory intent. A motor theory of speech perception, such as that advanced by the Haskins workers, is based on the observation that perception is often more closely related to articulation than to the acoustic signal. This view provides a framework within which the results of the present study receive a natural interpretation. If it can be said that listeners perceive consonants by reference to locus frequencies, then the subjects of this experiment perceived consonants by reference to loci appropriate to the vocal tract sizes producing the syllables they heard.

REFERENCES

- Darwin, C. (1971) Ear differences in the recall of fricatives and vowels. *Quart. J. of Exp. Psychol.* 23, 46-62.
- Delattre, P.C., A.M. Liberman, and F.S. Cooper. (1955) Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Amer.* 27, 769-773. (Also in Lehiste, 1967).
- Fourcin, A.J. (1968) Speech source inference. *IEEE Transactions* AU-16, 65-67.
- Ladefoged, P., and D.E. Broadbent (1957) Information conveyed by vowels. *J. Acoust. Soc. Amer.* 29, 98-104. (Also in Lehiste, 1967).
- Lehiste, I. (ed.). (1967) *Readings in Acoustic Phonetics*. Cambridge: MIT Press.
- Peterson, G.E., and H.L. Barney. (1952) Control methods used in a study of the vowels. *J. Acoust. Soc. Amer.* 24, 175-184. (Also in Lehiste, 1967).
- Stevens, K.N., and A.S. House. (1956) Studies of formant transitions using a vocal tract analog. *J. Acoust. Soc. Amer.* 28, 578-585.