SOME ACOUSTIC AND PERCEPTUAL CORRELATES OF SPEAKER IDENTIFICATION*

CONRAD LaRIVIERE

An investigation was undertaken concerning the ability of subjects to identify speakers solely on the basis of voice. The purposes of this study were: (1) to establish the relative contributions of source and vocal tract transfer characteristics to speaker identification, (2) to establish whether or not speakers could be identified on the basis of isolated utterances of continuant consonants, and (3) to investigate the nature of the relation between utterance intelligibility and speaker identification.

The subjects for this study consisted of eight male speakers and twelve listeners; the latter had been in routine contact with the former for a period of at least six months. The following speaker utterances, equated for intensity, were presented to the listeners: two prose sentences; four isolated vowels (/i, u, æ, a/) under three conditions; voiced, whispered and low-pass filtered at 200 Hz; and four isolated consonants (/s, f, v, z/).

The three vowel conditions were taken to SIMULATE the presence only of (1) source information (filtered vowels), (2) vocal tract transfer information (whispered vowels), or (3) both (voiced vowels). Except for the sentences, all stimuli were presented at a duration of 1250 msec. All stimuli were repeated five times and randomized.

The listeners were presented with forms listing each speaker by initials, and their task was to circle the speaker they felt produced each item. The listeners were also required to choose which stimulus item was presented for all the vowel and consonant stimuli employed in the study.

Acoustic analyses of the speakers' utterances were performed and the following parameters were extracted: fundamental frequency, the first three formant frequencies, the ratio of the second to the first formant frequency, formant bandwidths (for the voiceless consonants) and formant amplitudes (for the voiced and whispered vowels). The confusions among speakers predicted by each of these parameters were correlated with the actual confusions among speakers in an attempt to ascertain which acoustic characteristics serve as important cues to speaker identification.

The results of this study may be summarized as follows:

* Read by Howard B. Rothman.

(1) All stimuli yielded speaker identification performance at a level significantly above chance;

(2) The sentence stimuli resulted in performance (97 %) far above any other stimulus type.

(3) As shown in Figure 1, the performances achieved for whispered vowels (21.8 %) and filtered vowels (20.7 %) were very nearly equal, and, if summed, are close to the performance achieved for voiced vowels (40.2 %). Analysis of variance and *a posteriori* comparisons among means demonstrated, for voiced and whispered vowels, a general trend for low vowels to yield higher performances than high vowels. This trend is at least partially explained by acoustic analyses, which showed the *F2* and *F*3 formant amplitudes (*re*: *F*1 formant amplitudes) for the low vowels were considerably greater than those for the high vowels.

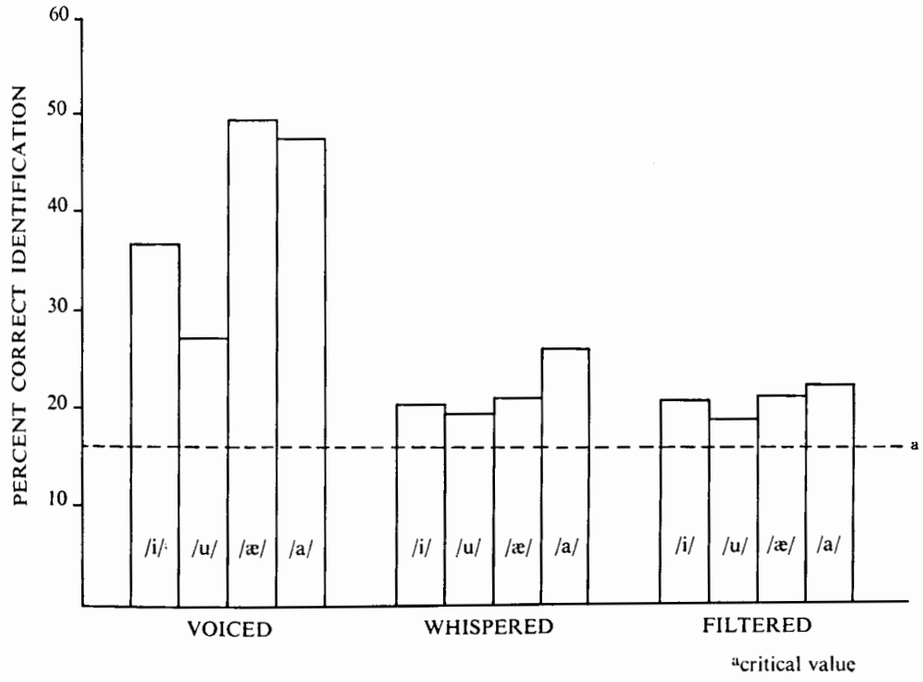

Fig. 1. Overall listener performance for vowel stimuli.

The correlations between acoustic characteristics and confusions among speakers (Figure 2) revealed that fundamental frequency, the second formant, and the third formant were, in general, equally good predictors of speaker confusions. This result seems to reinforce the notion that the contributions to speaker identification of the source and vocal tract transfer characteristics investigated here are equal and additive.

BASIS FOR Y RANKS

		X_1Y_1	X_2Y_2	X_3Y_3	X_4Y_4	X_5Y_5	X_6Y_6	X_7Y_7	X_8Y_8	ΣS	$\bar{S}$	$p(\bar{S})$
	f_o	8	20	13	13	4	13	13	3	87	10.9	.11
	$F1$	3	6	8	7	6	9	−1	−2	36	4.5	.32
/i/	$F2$	−3	3	5	−2	8	6	7	7	31	3.9	.36
	$F3$	11	5	10	6	7	6	7	2	53	6.6	.23
	$F2/F1$	−5	4	3	−4	7	8	6	4	23	2.9	. 4
	f_o	14	4	10	6	15	6	4	9	68	8.5	.18
	$F1$	4	0	−7	−5	3	3	−5	−3	−10	−1.25	.54
/u/	$F2$	3	5	16	10	17	10	4	14	79	9.9	.14
	$F3$	3	8	21	6	2	20	0	14	74	9.3	.15
	$F2/F1$	−2	6	17	13	16	15	1	9	75	9.4	.15
	f_o	14	6	0	3	6	5	3	6	43	5.4	.29
	$F1$	6	6	5	−3	4	7	10	−7	28	3.5	.38
/æ/	$F2$	−4	9	14	−4	11	9	15	9	59	7.4	.22
	$F3$	21	1	14	−2	9	4	7	0	54	6.8	.23
	$F2/F1$	1	8	9	−1	4	9	12	7	47	5.9	.27
	f_o	16	10	5	5	3	6	8	8	61	7.6	.22
	$F1$	8	7	12	3	0	1	2	9	42	5.3	.29
/a/	$F2$	11	−2	10	10	7	20	15	13	84	10.5	.12
	$F3$	9	18	10	4	9	10	7	12	79	9.9	.14
	$F2/F1$	4	11	7	11	5	6	−7	16	53	6.6	.23

Fig. 2. Rank Order Correlations Between Actual Confusions Among Speakers (Xi) and Expected Confusions Among Speakers (Yi) for Voiced Vowels.

(4) As shown in Figure 3, the voiced continuant consonants yielded significantly higher performances than their voiceless counterparts. Fundamental frequency was the best predictor of speaker confusions for the voiced consonants; for the voiceless consonants, the first formant frequency was the best such predictor obtained, though the correlation was weak in absolute terms.

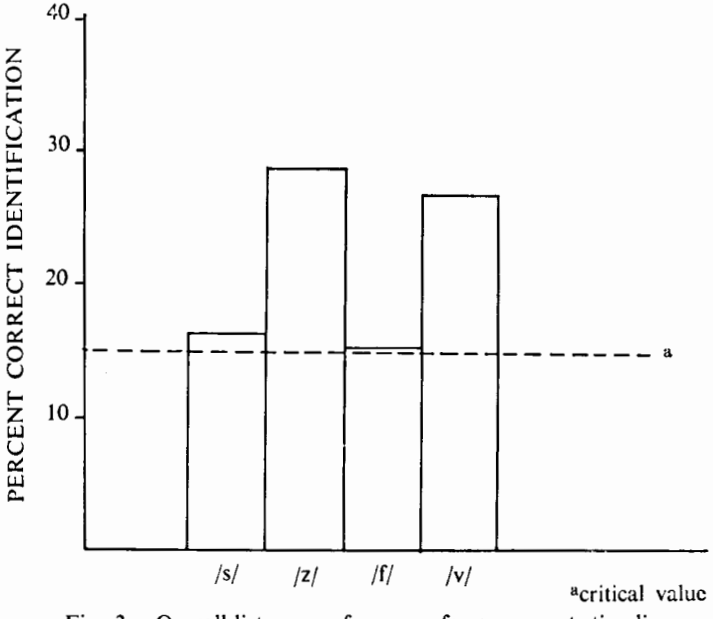

Fig. 3. Overall listener performance for consonant stimuli.

(5) The levels of utterance intelligibility are shown in Figure 4. Of interest here is the fact that, as a group, the filtered vowels were unintelligible; yet, it will be recalled that the speaker identification performance they yielded was very nearly that obtained for the highly intelligible whispered vowels. These results seem to indicate that utterance intelligibility is not a necessary concomitant to speaker identification. Figure 5 represents the distribution of response type by utterance. Note that by far the most common response type is "utterance correct, identification incorrect". This finding indicates that utterance intelligibility is not a sufficient concomitant to speaker identification.

It should also be noted that the acoustic parameters which have been traditionally associated with utterance intelligibility, (the $F1/F2$ ratio for vowels, and the second pole for consonants) do not correlate highly with speaker identification performance.

The major conclusions provided by this investigation are that, although one can point to acoustic correlates of speaker identification, there seem to be no acoustic invariants related to speaker identification; furthermore, speech intelligibility and speaker identification seem to be qualitatively different percepts. This would indicate that an adequate model for phoneme identification would not necessarily serve as

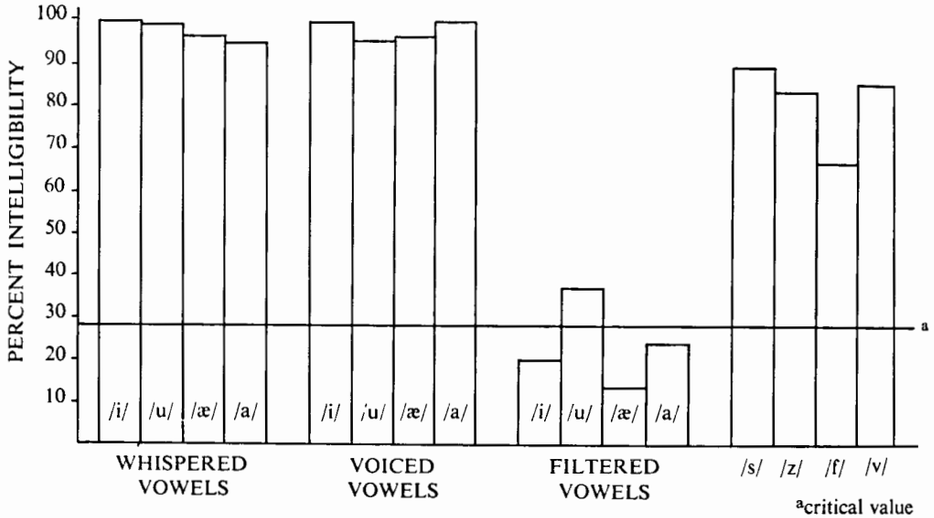

Fig. 4. Overall intelligibility levels of the utterances employed in speaker identity tasks.

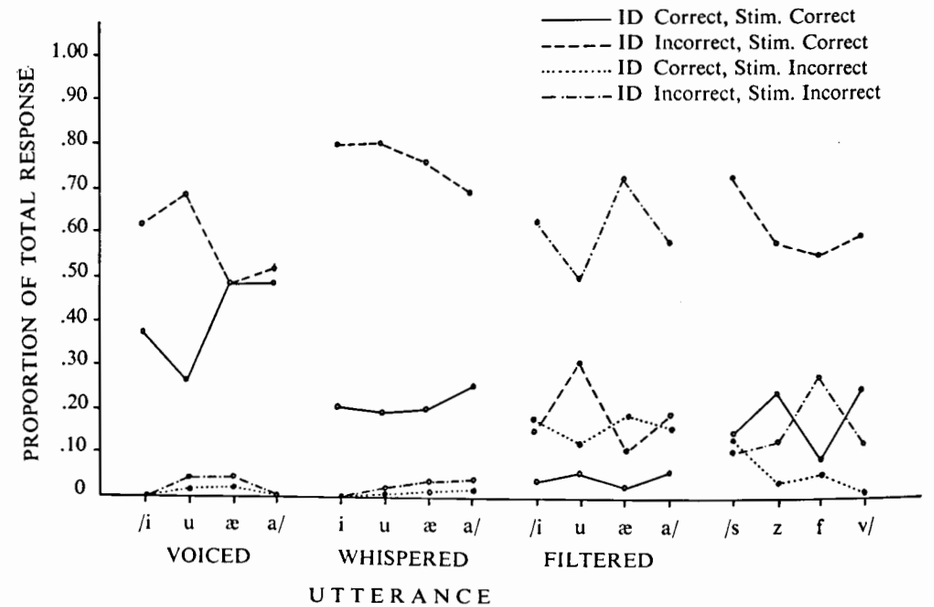

Fig. 5. Proportion of listener response types for utterances where intelligibility and identity judgments were made.

an adequate model for speaker identification and vice-versa. Further research into the nature and locus of speaker identification processing is strongly recommended, and a dichotic listening paradigm may prove particularly fruitful.

Finally, the very high performance obtained with the sentence stimuli points to the possible influence of supra-segmental features (such as tempo and inflection) in speaker identification. Research which attempts to isolate some of these factors is currently underway.

Speech Science Laboratory
University of Missouri
Kansas City, Missouri

DISCUSSION

FOURCIN (London)

Margaret Robertson, working at University College London (1971), has obtained results which are in general agreement with your findings but which are based on a more extreme range of speaker types, men, women, and children of the same dialect grouping, and with a different set of test items, only vowels in a consonant frame.

First, she found that speaker type identification, man, woman, or child, was not related to intelligibility. Second, however, she found that intelligibility was improved by giving a carrier phrase from the actual speaker to the listener before the test item was presented. Confusions were not reduced if the carrier was merely from the same speaker type.

I think it follows from this, as well as from other work (Fourcin 1968), that listeners can make use of precise information about a speech source, rather as though they were tuning in to its particular patterns. Listeners are not helped by a knowledge of approximate average *F2* or approximate vocal tract length.

REFERENCES

Fourcin, A.J.
1968 *IEEE Transaction on Audio and Electroacoustics Ref. AU* 16 (1968):65-67.
Robertson, M.A.
1971 "Some Effects of Source Inference on Speech Perception", Ph.D. thesis (University of London).

LaRIVIERE

The sources you site offer compelling evidence that listeners do engage in some sort of normalization process, based largely on source characteristics, as they are making speech recognition judgments. Indeed, the notion (Liberman *et al.* 1967) that vowels represent a cipher on the language, rather than a code, seems to demand that this sort of normalization occurs, in view of the large differences in vowel formant frequency values as a function of speaker type (Peterson and Barney 1952).

For the following reasons, however, I feel that one cannot at present resolve the issue of whether or not a similar normalization process contributes to speaker identification:

(1) Both Robertson's and the present work show no relation between speaker type or speaker identity judgments and speech intelligibility. There is then no *a priori* reason for inferring that cues pertinent to speech recognition are also pertinent to speaker identification.

(2) For the stimuli used here, there is no indication that listeners engaged in a speaker identification task more heavily weigh source characteristics than filter characteristics.

(3) Miller (1964) has presented evidence, based on inverse filtering techniques, which purports to show that the vocal tract transfer function carries more information about the identity of a speaker than does the glottal waveform.

In any event, I agree that the contributions of a normalization based on either source or filter characteristics constitute a viable research problem, and it should be relatively straightforward to devise suitable paradigms.

REFERENCES

Liberman, A.M., *et al.*
1967 "Perception of the Speech Code", *Psychological Review* 74:431-461.
Miller, J.E.
1964 "Decapitation and Recapitation, a Study of Voice Quality", *Journal of the Acoustical Society of America* 36:2002(A).
Peterson, G.E. and H.L. Barney
1952 "Control Methods Used in a Study of the Vowels", *Journal of the Acoustical Society of America* 24:174-184.