SPEAKER RECOGNITION FROM PHONATED VS. WHISPERED VOWELS UNDER DIFFERENT FILTERING CONDITIONS

WIM A. VAN DOMMELEN

Institut für Phonetik und
digitale Sprachverarbeitung
Universität Kiel
2300 Kiel, FRG

ABSTRACT

The perceptual contribution of glottal source and vocal tract characteristics to speaker recognition was investigated in two listening tests. A group of eight female speakers produced sustained /e/ and /o/ vowels in isolation, both whispered and phonated. 500 ms portions of these vowels were used as stimuli under different filtering conditions (0-1, 1-2, 2-5 and 0-5 kHz). The results indicate that neither these filtering conditions nor vowel quality exert systematic influence upon speaker identification. Glottal source information, however, proved to be of considerable perceptual importance.

INTRODUCTION

Relatively little is known about those perceptual cues in the acoustic speech signal that contribute to the recognition of speakers by the human listener. Following up investigations reported in the literature, e.g. /1, 2/, this paper examines the role of glottal source and vocal tract information via a direct comparison of speaker identification rates for phonated vs. whispered vowels. By taking isolated vowels spoken on a monotone speaker-specific supraglottal timing characteristics and pitch movements are ruled out as possible cues. At the same time, the question as to whether there are specific frequency domains of special importance was investigated by band-pass filtering and by the use of two vowels with different spectral composition.

PROCEDURE

A group of eight female German speakers (students of phonetics at Kiel University) produced sustained /e/ and /o/ vowels in isolation, both whispered and phonated. They were instructed to approximate vowel durations of about 1-2 seconds, a condition which was fulfilled with ease by all subjects. A second requirement concerned the phonated vowels, which had to be produced on a monotone. The pitch level, however, could be freely chosen at an individually comfortable level. Auditory examination as well as an analysis of the fundamental frequency showed that the vowels were produced with only slight perturbations, which were very unlikely to contribute to the recognition of the individual speakers.

After 5 kHz low-pass filtering (12 dB/octave) the speech material was digitized at a sampling rate of 10 kHz and manipulated in the following way: From the middle of each vowel a 500 ms portion was taken to serve as the raw material for a stimulus. Subsequently, loudness differences between the vowel portions were auditorily equalized by amplitude manipulation. After digital-to-analogue conversion the speech samples were filtered with three different bandpasses (0-1 kHz, 1-2 kHz and 2-5 kHz; 24 dB/octave) and re-digitized. Starting from this material, following a second digital-to-analogue conversion two stimulus tapes were constructed, containing whispered and phonated vowel portions respectively. Each tape comprised 64 stimuli (8 speakers x 2 vowel qualities x 4 filtering conditions; the fourth condition being 0-5 kHz) in a randomized order. Each stimulus was composed of a 100 ms warning tone and subsequent 0.5 sec pause followed by a vowel portion, repeated four times at intervals of 1.5 sec, and an ensueing 6 sec response pause.

The group of speakers, who were all well-acquainted with one another, also acted as listeners, being presented with the stimuli over a loudspeaker in a quiet room. The two experiments (whispered and phonated vowels respectively) were performed in two sessions separated by a week. The listeners indicated their answers by writing down the perceived identity of the speakers on a prepared answer sheet.

RESULTS

For clarity of presentation we will deal firstly with the effect of the different filtering conditions upon the speaker identification, secondly with the influence of vowel quality and, finally, with the role of glottal source. The statistical significance of the identification results was tested by means of Wilcoxon matched-pairs signed-rank tests.

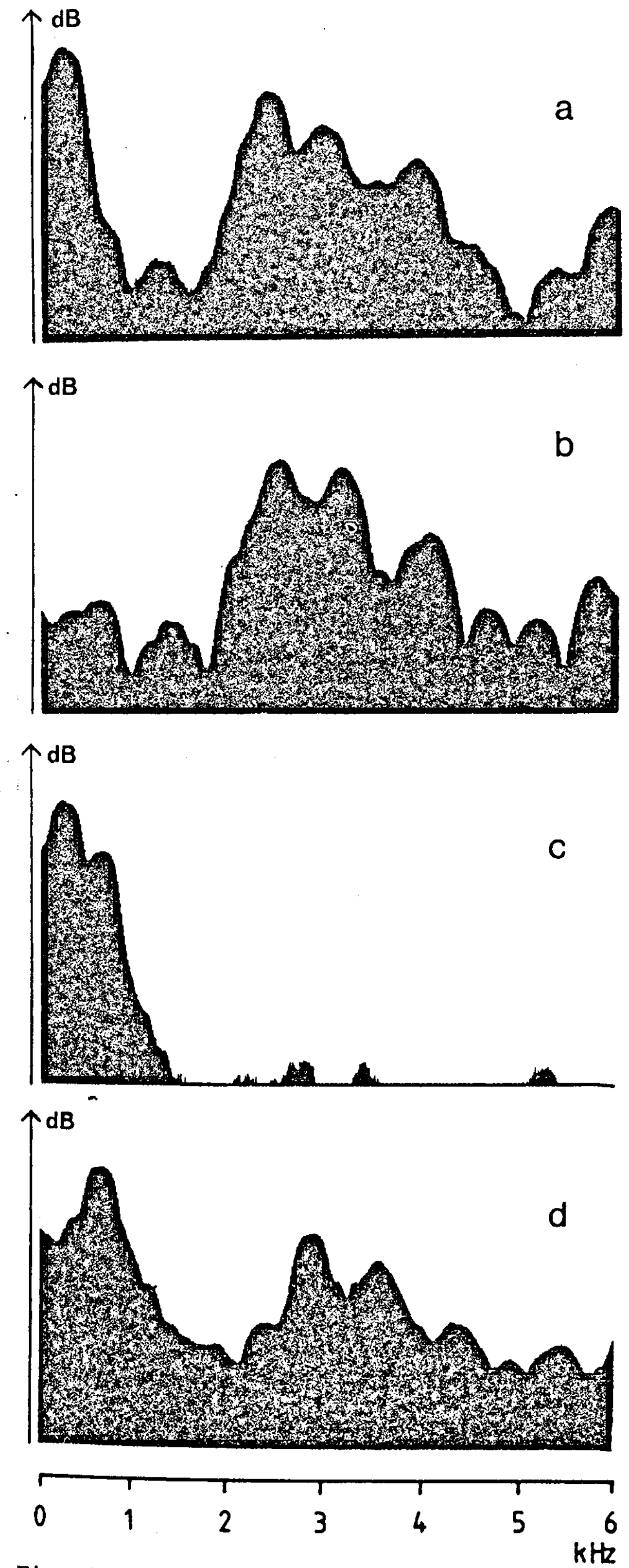

Fig. 1. Power spectra from four different vowels (speaker 2)
/e/: phonated (a) and whispered (b)
/o/: phonated (c) and whispered (d)

Filtering conditions

The overall spectra of the vowel portions are differently shaped (cf. Fig. 1), firstly due to the different positions of the first four or five formants in /e/ vs. /o/ (vowel quality) and, secondly, due to differences between the glottal spectra in phonated vs. whispered vowels (glottal source). Therefore, it might seem plausible to expect an effect of different filtering conditions upon the recognition rates (cf. Table I). Although in both cases there were sizeable differences between the best and the worst scores (16% and 13% respectively), they failed to reach statistical significance.

Table I
Overall identification scores
(percent correct)

Phonated vowels

0-1	1-2	2-5	0-5 kHz
29	35	35	45 %

Whispered vowels

0-1	1-2	2-5	0-5 kHz
13	21	10	23 %

Vowel quality

The overall recognition rates for /e/ vs. /o/ amounted to 34% vs. 38% in the case of phonated vowels and to 17% vs. 16% for the whispered vowels. In view of the small differences between the two conditions it is not surprising that they were statistically insignificant. This insignificance also holds when the individual filtering conditions are treated separately.

Glottal source

In contrast to both factors discussed above, the identification rates were affected by the glottal source parameter in a consistent way. At 36% the overall correct identification rate for phonated vowels lies significantly above that of 17% for the whispered ones (1% level). The effect holds for both vowel qualities (over all four filtering conditions) as well as for the filtering conditions (over the two vowel qualities), except for 1-2 kHz.

There are two aspects of the glottal source that might be responsible for the consistently higher identification scores in the case of the phonated vowels. In the first place, it is thinkable that speaker-specific pitch height was used as a primary cue in the identification task.

Alternatively, the spectrum of the glottal excitation is a possible candidate. With both possibilities in mind the data were examined further. Unlike the glottal spectrum, on which no data were available, mean fundamental frequency could be calculated for each speaker and turned out to vary between ca. 180 Hz and 250 Hz. Subsequently, a rank was given to each speaker, firstly according to their fundamental frequency (where a rank of 1 meant the speakers' own F0, a rank of 2 the next nearest pitch value etc.), and secondly according to their perceptual confusion with other speakers (over both vowel qualities and all filtering conditions; a rank of 1 standing for the highest recognition rate etc.). Calculation of Kendall correlation coefficients showed significant relationships (for one speaker at the 5% level; otherwise 1%) with values of r= 0.57, 0.67, 0.69, 0.72, 0.73, 0.76, 0.84 and 0.96. Obviously, speakers with similar fundamental frequencies are far more likely to be confused than speakers showing different pitch height. So it seems that the listeners relied upon the F0 factor to a varying, sometimes rather high degree in their identification of the various speakers.

Following a procedure similar to the one described above, correlations between perceptual confusions in the phonated vs. the whispered condition were calculated. Since the information present in whispered vowels is almost exclusively vocal tract information, it was postulated that high correlation rates might indicate a high perceptual value of such information. These correlation rates turned out to be significant for only two speakers (r= 0.64 and 0.81 respectively, at the 5% and 1% level respectively). Overall recognition rates for these speakers happened to be the highest ones (48% and 47% respectively for phonated vowels as against 34% and 28% for whispered vowels). Therefore, it seems likely that vocal tract information served as a perceptual cue in these two cases in addition to the glottal source parameter.

DISCUSSION

Of the three factors investigated in this paper, various filtering conditions, vowel quality and glottal source, only the latter turned out to have a systematic influence upon the speaker identification scores. The enhancing effect of glottal source information on identification can probably be accounted for by the speaker-specific pitch height, which is in line with the findings of Compton /1/.

Further, the results suggested a predominance of the glottal source parameter over vocal tract filtering characteristics. This confirms the findings of Lass, Hughes, Bowyer, Waters and Bourne /3/ for speaker sex identification. Possibly due to the use of synthetic stimuli instead of natural speech Lehiste and Meltzer /4/ arrived at the opposite conclusion, whilst LaRiviere /2/ found both factors to contribute about equally to speaker recognition. One should note, however, that in the present paper no data about the contribution of the glottal spectrum was available, so that its minor relevance had to be inferred from the data on fundamental frequency and from the mostly weak correlations between perceptual confusions for phonated vs. whispered vowels.

The fact that the four filtering conditions failed to influence the listeners' identification judgements may be due to there being only 16 stimuli in the sample (8 speakers x 2 vowel qualities per filtering condition; cf. the clearer effect of 24% for 1020 Hz low-pass vs. 1020 Hz high-pass found by Compton /1/ using considerably more stimuli). However, with vowel quality the sample size (32) is twice as big (8 speakers x 4 filtering conditions per vowel quality); this increases the likelyhood of the vowel quality results being representative.

REFERENCES

/1/ A. Compton, "Effects of filtering and vocal duration upon the identification of speakers, Aurally", Journ. of the Acoustical Society of America 35, pp. 1748-1752, 1963.

/2/ C. LaRiviere, "Some acoustic and perceptual correlates of speaker identification", Proc. of the Seventh Int. Congr. of Phon. Sci. (Rigault, A.; Charbonneau, R., eds.); Mouton: The Hague/Paris, pp. 558-564, 1972.

/3/ N.J. Lass, K.R. Hughes, M.D. Bowyer, L.T. Waters, & V.T. Bourne, "Speaker sex identification from voiced, whispered, and filtered isolated vowels", Journ. of the Acoustical Society of America 59, pp. 675-678, 1976.

/4/ I. Lehiste, & D. Meltzer, "Vowel and speaker identification in natural and synthetic speech", Language and Speech 16, pp. 356-364, 1973.