

Telephone speaker recognition amongst members of a close social network¹

Paul Foulkes and Anthony Barron***

**Department of Language and Linguistic Science
University of York*

***Department of Linguistics and Phonetics
University of Leeds*

ABSTRACT This article presents results from a speaker recognition task carried out by a close-knit network of speakers (university friends who have lived in shared accommodation with each other for two years). Ten male speakers recorded a scripted message onto an answer machine via a mobile telephone. Two foil speakers from outside the network were also recorded. Samples of between 8 and 10 seconds were extracted from all twelve recordings, and used as stimuli for an open speaker recognition test performed by the network members. Listeners varied widely in their performance, and one listener failed to recognize his own voice. Some of the voices were easy to identify, but several speakers were consistently misidentified, and one speaker was particularly hard to identify. Both of the foil speakers were sometimes mistaken for network members. Auditory analysis of the voices shows, as expected, that speakers with the most distinctive regional accents and other idiosyncratic features were the most consistently identified. Acoustic analysis of F0 was also undertaken. It was found that the speakers who were most consistently identified were those with relatively high and low mean F0 values, as well as those with the widest and narrowest overall F0 range. Speakers with average pitch values and ranges in the middle of the overall group values proved harder to identify. The findings support the view that average pitch is a robust diagnostic of speaker identity, not only for forensic phoneticians, but also for naïve listeners. They furthermore demonstrate that naïve speaker recognition, even among members of a close-knit social network, is not a task which can be achieved infallibly.

KEYWORDS speaker recognition, speaker identification, telephone speech, fundamental frequency, English

INTRODUCTION: SPEAKER RECOGNITION

Identifying the speaker responsible for producing a particular sample of speech is probably the most common aim in forensic linguistic analysis. This process of speaker recognition (henceforth SR) has been defined by Nolan (1997: 744) as 'any activity whereby a speech sample is attributed to a person on the basis of its phonetic-acoustic or perceptual properties'. Depending on the circumstances of the legal case, SR may be performed by professional linguists, and/or by untrained, naïve listeners. Where speech analysis by trained linguists is concerned, it has been estimated that upwards of 95 per cent of cases involve analysis of recorded telephone

calls (Künzel 1997). Given the relative frequency of analysis of this type, much research has been carried out to identify the phonetic parameters by which one voice may differ from another. Some of the more robust diagnostics include voice quality, the regularity and phonetic quality of hesitation markers, and average pitch (see further e.g. French 1994, Nolan 1991, 1997).

Naïve speaker recognition is an activity which all human beings undertake in everyday situations. Perhaps the most obvious example of this is when a person answers a telephone call: even though the caller's face cannot be seen, and the telephone transmission causes the acoustic signal to undergo various modifications and degradations, the listener can usually identify a familiar voice within a few seconds. The commonest forensic scenario in which naïve SR comes into play again involves telephone interaction, for instance when the crime involves the sending of abusive or threatening calls. If the calls have not been recorded, the witness may be asked to attempt an identification of the voice through a voice line-up, in which recordings are played of the suspect's voice along with the voices of several foils (see Künzel 1994, Hollien 1996, Nolan and Grabe 1996).

The cognitive processes by which naïve SR takes place are not well understood. Obviously, SR must be effected with reference to linguistic cues, which may be lexical, syntactic, morphological and pragmatic as well as phonetic and phonological. However, it is not clear precisely which linguistic/phonetic cues are identified by the listener, and nor is it clear what the relative importance of such cues is with respect to other cues. As likely as not, the relevant cues and their respective weighting will differ from situation to situation. It is furthermore evident from anyone's experience of everyday SR that the cognitive processes can sometimes fail. For example, everyone who receives phone calls makes the occasional mistake in recognition, either by failing to identify a known caller, or by wrongly identifying the caller as someone else.

SPEAKER RECOGNITION OF TELEPHONE SPEECH

Various reasons can be identified to explain why SR is not a straightforward task. With telephone speech in particular, problems occur because some of the linguistic and phonetic characteristics of a voice may be altered by the telephone context itself. Some of these differences are caused by the technical effects of transmission. For instance, sound frequencies below c. 300 Hz and above c. 3,400 Hz are removed, such that much of the essential speaker-specific information encoded in the third and fourth formants of vowels may be missing (Nolan 1983, Künzel 1995, 1997). The context in which calls are made may also introduce a high degree of background noise, for example pay phones used by busy roads, or mobile phones used in public places.² Background noise masks the acoustic information of the intended speech signal. Secondly, some differ-

ences between telephone and direct speech appear to be induced by speakers themselves. Several studies have shown that speakers on the telephone tend to speak louder and raise their fundamental frequency, and thus the perceived pitch of their voices (e.g. Hirson *et al.* 1994; Braun 1995; but see Künzel 1997 for conflicting evidence). It is also well known that some speakers adopt a 'telephone voice', modifying their rate of speech, segmental pronunciations, and/or voice quality (Wells 1982: 28).

The sum of differences between direct and telephone speech support the view that individuals need a certain amount of time and experience in order to develop what Baldwin and French (1990: 110) have termed *telephone recognition strategies*: that is, specific cognitive mechanisms for identifying voices in the telephone context. An important implication of this for forensic linguistics is that familiarity with a voice in face-to-face interaction, and successful SR of that voice in a direct setting, do not reliably demonstrate that the same voice will be equally well recognized if the context is changed to telephone interaction (Baldwin and French 1990: 111). It also follows from this observation that the success rate of any type of SR depends to some extent on the familiarity of the voice concerned, such that less familiar voices may prove more difficult to identify (Stevens *et al.* 1968, Hollien 1990, 1996).

It seems we must also acknowledge the 'rather gloomy' conclusion reached by Hollien (1996: 15), that certain individuals are simply not very good at SR, be it over the phone or in direct contact. Evidence to this effect is provided by Shirt (1984), who carried out several SR tests with groups of phoneticians and naïve subjects. It emerged that individuals of both groups varied widely in their performance, with success rates varying from 76 per cent to 38 per cent amongst phoneticians, and from 76 per cent to 19 per cent within the lay group. More worryingly still, Shirt also found that phoneticians performed only marginally better than the untrained respondents.

Finally, it is important to note that these facts about the variability of SR conflict with assumptions commonly held by members of the legal profession, namely that recognition of voices is an easy and uncomplicated task, especially when the voice belongs to an individual who is well known to the listener (Künzel 1995; McClelland 2000).

Research into SR has typically focussed on performance by listeners presented with direct speech. Rather little research has so far been undertaken to assess the reliability of naïve listeners in SR performance with telephone speech (but see Künzel 1990). A striking exception is the study presented by McClelland (2000), who devised a closed SR experiment involving fourteen male and female members of three generations of her immediate family. Each individual was recorded making two phone calls. The first call involved the reading of a short story via a land-line. The second call consisted of a scripted message using a mobile phone, the role of the script being to control for any non-phonetic differences between speakers. McClelland extracted

short samples (approximately forty words) from each of the recordings and from them compiled a test tape for a SR task. Each of the family members listened to the tape once, with the instruction to identify which family member was responsible for the sample. The listeners were explicitly told to select their answers from the possible set of fourteen speakers. In spite of the anticipation that the task would be easy, the results proved surprising. Some voices were indeed recognized with no problem, but these were generally voices which stood out from the group as a result of their marked phonetic characteristics (for example, the speaker with the lowest pitch, and a speaker with a regional accent different from that of the other members). Some listeners proved poor at the task, making incorrect judgements or failing to make any decision as to which of the family group had produced the sample. Some listeners even denied that certain samples were produced by any of the family members.

In the remainder of this article we describe another SR experiment which focuses on telephone speech. This experiment is modelled on McClelland's study, but is characterized by various differences in the experimental design.

EXPERIMENT

Like McClelland (2000), our study assesses SR by a group of people who know each other very well. Our group, however, consists of a set of young men who are university friends. This group was selected to investigate SR in a situation where the social profile of all group members is very similar in terms of age and gender (compare with McClelland's study, which involved men and women of various ages). Our SR task also differs from McClelland's by virtue of being open rather than closed. That is, the listeners were not assisted in making their identification by being told who the possible speakers were. Such a scenario more closely mimics real-life forensic settings. To this end we also included foil voices from speakers outside the main group to assess their impact on the SR task.

Selection of subjects

The main subjects used in this experiment (henceforth referred to as the network members) were ten second-year undergraduates of Leeds University. The network members included the second author, who acted as the experimenter. All were male, aged twenty or twenty-one, and formed a close social network. During their first year as students the ten had all lived together in shared student accommodation. Some of the network members spent large proportions of their academic time together, and they had all socialized with each other on a regular basis. In their second year, at the time of the experiment, the ten lived in three separate but geographically close houses. Four of the network lived in one house,

five shared another, and the tenth member lived with a different set of friends. (In the discussion of the results below, we refer to these as the three subsets of the main network.) All ten still socialized with each other on a regular basis, and by this time had known each other well for a period of approximately twenty-one months.

As is to be expected, the subjects were of varying geographical origin and hence display some differences in their regional accents. One subject, Alex, came from Tyneside, an area well known for its distinctive accent. Another subject, Rich, had a relatively strong, non-standard London accent. The other subjects' accents, however, showed close similarities with each other (as judged by both authors). The most salient features of these accents can be considered general for young speakers from the south of England. These include several non-standard features, such as the regular use of [ʔ] for phrase-final and intervocalic /t/, and vocalization of coda /l/. Aside from accent features, the only other obviously significant feature of the voices used is the fact that one speaker, Pete, had an occasional stammer. We return to consider these issues in the discussion of the results below.

Two other male subjects were also used in the investigation, to act as foils in the SR test. These two, who included the first author, were unknown to any of the network members (with the obvious exception of the second author). The sociolinguistic characteristics of the two foils were similar to those of the majority of the main group.

Selection of language material

Following McClelland (2000), each of the twelve speakers was recorded sending a scripted message via a mobile telephone. The scripted message was used to control for any linguistic variation beyond the phonetic/phonological level. The full script is given below:

Hi Anthony, X here. Just calling to check how you are and how the rest of the house are getting on with their revision. All our exams finish a week on Wednesday.

Anyway, there are two things I need to tell you about ...

Firstly – I may be getting a job somewhere down near your house this summer: do you think you'll be around? If not, do your parents have a comfy sofa up for grabs?

Secondly – and most importantly – some of the boys from home are having a bit of a get-together on Friday. This should involve a barbecue and fair amount of beer... let me know if you'll be free for that. That's Friday from 6-ish.

Right then, that's it. Give us a call soon at the house on 01234 567 890.

See you/bye/catch you later [etc.]

For the purposes of eliciting speech that was as natural as possible, the subjects were asked to rehearse reading the script, and then to deliver it as if it were a spontaneous message to the second author. The subjects were instructed not to insert words which were not included in the written script. However, the closing line was left vague on purpose, since it is well known that telephone closures differ from person to person (Laver 1981), and it was not our intention to use the full text in the listening test. A Bosch 909s dual band telephone was used for the message elicitation, and the spoken script was recorded via a Betacom Solo answer machine onto a Sanyo microcassette.

Sampling

The twelve answer machine messages were transferred onto a TDK AR46 (type 1) compact cassette tape via a Technics stereo double cassette deck. A brief auditory analysis of the samples helped to determine which part was to be used in the listening test. In spite of instructions to the contrary, some speakers did modify the script of the message. One speaker failed to read the telephone number at the end of the message preferring 'you know the number', while another inserted several tags and fillers such as 'as you know' and 'you know?'. In order to exclude sections of the message which contained unscripted material, the following forty-two-word section was selected:

there are two things I need to tell you about ...

Firstly – I may be getting a job somewhere down near your house this summer: do you think you'll be around? If not, do your parents have a comfy sofa up for grabs?

The only lexical variation found in these sections were the omission of the word *sofa* by John, and of the phrase *up for grabs* by Rob. The selection of forty-two words was similar to that used by McClelland (2000), and the average sample length was 9.2 seconds. The twelve samples were compiled in random order to create a new tape for the purposes of the listening test. A gap of 5 seconds was left between samples on the test tape.

The listening test

The listening test was administered to nine of the network members (the exception being the second author, who administered the test). The listeners were presented with a simple questionnaire and were asked to write down the names of the speakers on the tape in the order they heard them. Unlike McClelland (2000), who informed her listeners who the possible speakers were, our test was an open one. The respondents were told that

they might or might not know the identity of the speakers on the tape. It was also pointed out that the same voice might be heard more than once (although in practice none of the voices was included twice). The tape was played only once to each listener.

Results

The results of the listening test are presented in full in Table 1. The twelve speakers are listed on the left, while the nine respondents are listed at the head of each column. The network members are divided into three subsets, referring to the three different houses in which they lived. Correct responses are indicated with a tick, ✓. Where an incorrect identification was made, the name given by the respondent is listed. Empty cells indicate no response was given.

Table 1 Responses to speaker identification test

Speakers ↓	Listeners								
	Set 1	Set 2					Set 3		
	Bill	John	Andy	Hugh	Ben	Steve	Rich	Alex	Pete
Set 1 Bill	✓	✓	✓	✓	✓		✓	✓	
John	Andy	✓	✓	✓		✓	✓		✓
Andy	Ben	Steve		✓		✓			Ben
Set 2 Hugh	✓	Pete		✓			✓	✓	✓
Ben		✓		✓	✓	✓	✓	✓	Steve
Steve	Rich	✓		✓		✓			✓
Rich	✓	✓	✓	✓	✓	✓	✓	✓	✓
Set 3 Alex	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pete	✓	✓	✓	✓	✓	✓	✓	✓	✓
Anthony	✓	Andy		Hugh	✓		✓	✓	✓
Foils Rob		Ben							
Paul	Hugh						Hugh		

Table 1 shows quite clearly that listeners displayed varying levels of ability in the speaker identification task, and also that some speakers were more readily identified than others. We discuss first the performance by the individual listeners, before turning to an analysis of the speakers.

Results by listener

The bottom two rows of Table 1 indicate that the two foil voices were wrongly attributed to network members on three occasions. John incorrectly identified Rob's voice as that of his own housemate, Ben. Two other listeners wrongly identified Paul's voice, and in both cases attributed it to Hugh.

Table 2 Summary of responses by listeners to samples drawn from network members

Listener	Correct	Unidentified	Incorrect
Bill	6	1	3
John	7	-	3
Andy	5	5	-
Hugh	9	-	1
Ben	6	4	-
Steve	7	3	-
Rich	8	2	-
Alex	6	4	-
Pete	7	1	2
total (N=90)	61	20	9
percentage	67.8	22.2	10.0

Table 2 summarizes the performance of each of the nine listeners to the ten network voices. The scores for the entire group of listeners are also included in the bottom two rows. Note first that the overall success rate for the listeners was 67.8 per cent (that is, 61 of 90 answers were correct). More than one in five of the experimental stimuli went unidentified, and there were nine incorrect identifications, which amounts to a 10 per cent error rate in given answers by the group as a whole.

No individual listener managed to identify all of the network voices correctly. The best performer was Hugh, who correctly identified nine voices. However, he mistakenly attributed Anthony's voice to himself. Four of the listeners made at least one incorrect identification, with both Bill and John making three errors. Seven of the nine listeners were unable to make any identification of at least one stimulus. Andy, Ben and Alex had the highest rate of non-identification, with four or five voices unattributed. Andy, in fact, was also the only respondent who failed to identify his own voice.

Also revealing is a closer examination of the results by speaker subset (which are divided according to the accommodation arrangements of the network members). The leftmost response column of Table 1 shows that Bill, the listener who lives apart from the others, correctly identified all members of Set 3, but only one member of Set 2. The members of Set 2 perform surprisingly poorly when it comes to identifying each other, making only fourteen correct responses from a possible twenty-five. Hugh's voice, for example, is correctly identified only by Hugh himself amongst the Set 2 listeners. Ben manages to identify himself but none of his housemates, while John makes two errors of judgement. The members of Set 3 appear to perform rather better in identifying each other, making eleven out of twelve correct responses (the only flaw is Alex's lack of response to Pete's voice).

However, as we shall discuss in the next section, the voices of Set 3 were in several respects the easiest to identify. Note also that the members of Sets 1 and 2 also perform well in general with the Set 3 stimuli.

Results by speaker

The responses to the stimulus voices are summarized in Table 3. This Table indicates the overall number of correct and incorrect responses, and the number of times a voice was not identified.

The results are similar in kind to those obtained by McClelland (2000), in that some voices posed few problems to listeners, while others proved much more difficult. Two voices (Rich and Alex) were correctly identified by all nine listeners, while another (Pete) was rightly named eight times. The response rate to the other voices varies substantially, however. The least frequently identified voice is Andy's. Not only is he misidentified three times, as two other members of the network, he is only correctly identified twice. Moreover, as was pointed out in the last section, he is the only subject who could not identify his own voice. Interestingly, a relatively poor success rate was also found with Anthony's voice, even though he was actually administering the test to the others. The foil voice of the first author was incorrectly attributed to Hugh by two listeners. No names from outside the network were offered by any of the listeners, which perhaps bears witness to the closeness of the social group.

In the next section we analyse the voices in more detail, both auditorily and acoustically, in order to offer some explanations as to why certain voices prove easier to identify than others. Although we obviously focus

Table 3 Summary of responses to each speaker sample

Speaker	Correct	Unidentified	Incorrect	Mistaken for
Bill	7	2	-	-
John	6	2	1	Andy
Andy	2	4	3	Ben (x2), Steve
Hugh	5	3	1	Pete
Ben	6	2	1	Steve
Steve	4	4	1	Rich
Rich	9	-	-	-
Alex	9	-	-	-
Pete	8	1	-	-
Anthony	5	2	2	Andy, Hugh
Rob (foil)	(n/a)	8	1	Ben
Paul (foil)	(n/a)	7	2	Hugh (x2)

on the voices of our network members in this specific test, we raise some points which we suggest may be of value in understanding the processes of naïve SR in general.

ANALYSIS OF SPEAKER VARIABLES

Auditory analysis

Following the execution of the SR test, both authors undertook close auditory analysis of the samples to identify the most salient phonetic characteristics of each voice. We also interviewed two of the speaker-listeners, Alex and Rich, and played the samples to them again, in order to gauge which phonetic parameters they were consciously aware of.

The easiest question to answer is why Rich, Alex and Pete should be the most consistently identified. This finding is predictable on the basis of their relatively idiosyncratic speech characteristics, as outlined above in the section on speaker selection. Rich and Alex have the strongest regional accents among the speakers (London and Tyneside respectively), while Pete is the speaker with a mild stammer. In the speech samples used in the SR test, all three displayed phonetic and phonological features indicative of these factors. Alex, for example, displayed his northern roots with the use of [ʊ] in *us* and *up*. He specifically cued his Tyneside origin by producing glottalized stops in *put us*, and *need to*, and by producing non-glottalized stops in pre-pausal positions (*about...*, *if not...*), a feature which differentiates him from all other speakers in the sample (see further Wells 1982, Docherty and Foulkes 1999, Watt and Milroy 1999). Similarly, Rich's strong London accent is revealed by phonetic forms such as his very fronted [a]-like pronunciation of /ʌ/ in *somewhere*, *summer* (Wells 1982, Tollfree 1999). Pete, meanwhile, produced two short stammers in his sample, repeating the syllable [tə] in the phrase *to tell you about*, and prolonging the initial [s] of *somewhere*. Note that Pete, Rich and Alex are all members of Set 3 of the network, which explains why the SR results for this Set were so much better than those for Set 2 (Table 1).

When interviewed about the samples, both Rich and Alex were able to describe these features of these three voices, indicating their overt salience as cues for the SR task. However, it should be borne in mind that even in spite of his stammer, Pete was not recognized by his housemate Alex in the SR test itself! And by contrast, one listener wrongly attributed Hugh's voice to Pete.

Analysis of the general segmental and suprasegmental characteristics of the voices also helps explain some of the problems and errors that occurred in the SR test. As already noted, the accent features of the speakers other than Rich, Alex and Pete were generally very similar to each other, with various non-standard consonantal variants striking both authors as salient in most or all of the samples. It is therefore less surprising to find uncertainty

or confusion with respect to these other voices than those of the three speakers with very marked individual characteristics.

Close auditory analysis of the samples did, however, reveal that the voices which were less well identified none the less contained phonetic cues which were not found in some or any of the other samples. This suggests that some cues, however salient they may appear to phoneticians, are not particularly useful diagnostics in the process of 'live' speaker recognition. Andy, for instance, uses a diphthong in the first vowel of *sofa* which is characteristic of his south-western origins, and which differs quite substantially from the vowel used by all other speakers. Nevertheless, Andy's voice is the least successfully identified of the whole network. Steve and Ben both displayed a markedly creaky phonation quality which was not found in the other voices. While it is worthy of note that Ben's voice was attributed to Steve by one listener, it is even more striking that Steve's voice was only identified correctly four times, making it the second most difficult. Furthermore, two of the non-creaky voices were wrongly credited to Steve. These findings suggest that phonation type was of limited relevance for this SR test.

When commenting on the samples, both Alex and Rich repeatedly made reference to the speakers' pitch. Three speakers (John, Ben and Rich) made use of the high rising terminal intonation pattern (or 'uptalk'; Cruttenden 1997: 129ff.). This involves rising intonation contours in non-sentence final positions, apparently used as a strategy for monitoring a listener's attention or understanding. It is furthermore an innovative and highly salient pattern, to the effect that it merits (invariably negative) comment in newspaper articles and letters (see for example Bradbury 1996). Although the presence of this intonation pattern was overtly commented upon in our post-test interview, it, too, appears to have had little impact on listeners in the test itself. John and Ben were identified by only six of the nine listeners, while both Ben (three times) and Rich were wrongly identified as other speakers who did not display the pattern.³

Other comments about the samples referred to the general pitch level. For example, Pete's voice was described as 'deep' and 'constant', Hugh's was 'high', and both Hugh and Paul were felt to have 'extreme' or 'up-and-down' pitch contours. In the light of such descriptions, coupled with the fact that fundamental frequency is invariably measured by phoneticians in forensic speaker identification cases, we performed acoustic analysis of each speaker's F0 to assess its possible contribution to the SR test. The results are presented in the next section.

Acoustic analysis of fundamental frequency

Acoustic analysis of the samples was carried out using both Sensimetrics SpeechStation 2 and Praat software. The samples from the test tape were first digitized at 11,025 Hz via the Sensimetrics software. The resultant

sound files were then analysed using the Praat system, which provides a more sophisticated analysis of average F0. The parameters for analysis were set to 80 and 200 Hz, since this is the normal range of F0 for an adult male speaker (e.g. Ladefoged 1993: 187). Praat calculates for the sample selected a maximum and minimum F0 value, as well as the overall mean F0 and standard deviation from the mean. The standard deviation is a useful measure for our purposes, since it offers a quantification of the degree of F0 movement a speaker uses. More monotonous voices are reflected by a lower standard deviation. We present the results of this analysis in Table 4, which also repeats the number of correct identifications of each voice from Table 3. The average scores for all ten network members are also shown in Table 4.

The average F0 of all the speakers, 120.9 Hz (which increases to 123.6 Hz if the foils are also included) is comparable with that of Künzel (1997), who found a mean of 129 Hz for a group of German males. It is considerably higher than that found by Hirson, *et al.* (1994), however, whose twenty informants produced a group mean of 108.9 Hz in unscripted telephone speech. We are unable to discuss further the possible implications of our data in this respect, since we do not have recordings of our speakers in non-telephone settings for comparison.

Two of the most conspicuous false judgements in the SR test may be in part explained with reference to these F0 measurements. First, recall from Tables 1 and 3 that Andy is twice mistaken for Ben. We have already

Table 4 Average F0 values and standard deviations

Speaker	Mean F0 (Hz)	s.d. (Hz)	Correct ID (max = 9)
Bill	127.3	23.5	7
John	138.7	17.8	6
Andy	120.8	17.2	2
Hugh	129.9	22.5	5
Ben	125.0	15.4	6
Steve	106.2	15.2	4
Rich	150.4	22.7	9
Alex	105.8	10.2	9
Pete	92.5	9.5	8
Anthony	112.4	13.9	5
<i>all network</i>	120.9	16.8	6.78
Rob (foil)	135.7	17.5	0
Paul (foil)	138.8	30.7	0

explained that their general accent features are similar to each other, but note also that the average F0 of these two speakers differs by less than 5 Hz, and that their standard deviations differ by less than 2 Hz. Similarly, the foil voice of the first author is twice attributed to Hugh. Again, there were few obvious segmental differences between their voices, but it may also be of relevance that both are notable for their high standard deviations, which indicate a relatively wide pitch excursion over the course of their samples. (Recall that their high degree of pitch movement was explicitly referred to by the subjects we interviewed.)

However, further examination of F0 measurements shows that some of the SR errors involve pairs of speakers with widely different F0 values. This is most notable in the attribution of Hugh's voice to Pete. Hugh has a higher than average F0, while Pete's is by far the lowest of the whole network (and some 37 Hz lower than Hugh's). Moreover, although Hugh has a high standard deviation of 22.5 Hz, Pete has the lowest standard deviation, indicating that his is the most monotonous voice among the samples. Thus it appears that where errors are made in the SR task, even very marked differences in the speakers' F0 may be ignored by listeners.

On the other hand, if we concentrate on the correct responses, rather than on the errors, the data in Table 4 do in fact suggest that F0 values may in general play a positive role in speaker recognition. The speakers who are most often correctly identified are usually those with the most extreme F0 values of the group. Pete and Alex, notwithstanding their otherwise recognizable idiolects, also have the two lowest mean F0 values. Rich, by contrast, has the highest mean F0. Rich and Bill both register high standard deviation scores, while Pete, as we have seen, has the lowest.

Table 5 Difference of F0 values and standard deviations from the network average (mean = 120.9 Hz, s.d. = 16.8 Hz)

Speaker	Δ mean F0 (Hz)	Δ s.d. (Hz)	Correct ID (max = 9)
Bill	6.4	6.7	7
John	17.8	1.0	6
Andy	0.1	0.4	2
Hugh	9.0	5.7	5
Ben	4.1	1.4	6
Steve	14.7	1.6	4
Rich	29.5	5.9	9
Alex	15.1	6.6	9
Pete	28.4	7.4	8
Anthony	8.5	2.9	5

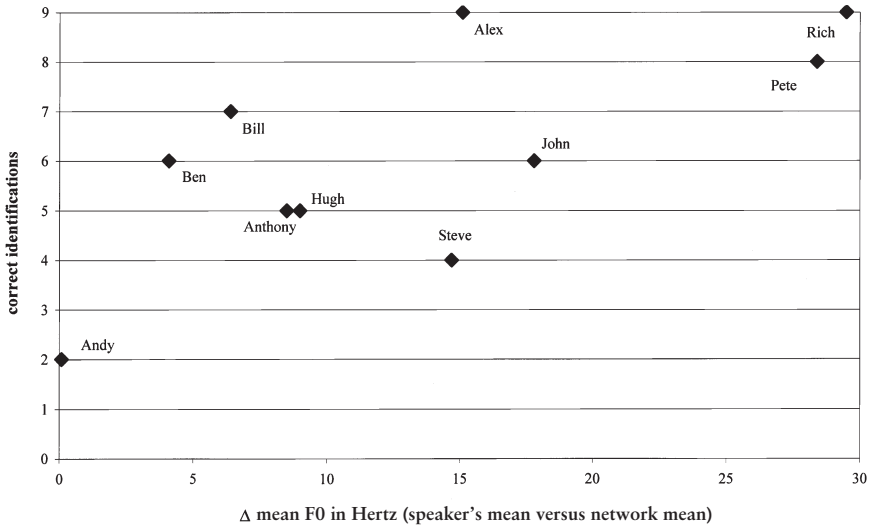


Figure 1 SR results as a function of mean F0 (the difference between speaker's Δ mean F0 and the overall network mean)

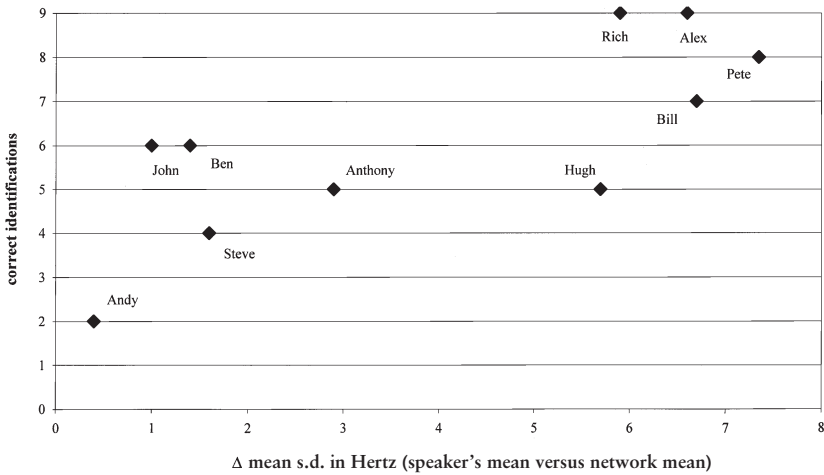


Figure 2 SR results as a function of Δ s.d. of F0 (the difference between the speaker's s.d. of F0 and the overall network mean)

Speakers who have less peripheral scores, closer to the group means, tend to be the ones less frequently identified. See, for example, the scores for Andy, Steve and Anthony.

In order to test whether average F0 and/or standard deviation of F0 make any statistically significant contribution to the recognition test, we have recalculated the data from Table 4 to reflect the degree of difference of each speaker's values from those of the group as a whole. These recalculated figures are shown in Table 5, and graphically in Figures 1 (mean F0) and 2 (standard deviation).

The figures in Table 5 show the absolute difference between the individual scores and the overall group scores given in Table 4 (mean = 120.9 Hz, s.d. = 16.8 Hz). Thus, for example, Andy's mean F0 is 120.8 Hz compared with the network mean of 120.9 Hz. This difference of 0.1 Hz is the Δ mean value displayed in Table 5. It is in fact the lowest Δ mean score in Table 5, indicating that Andy's average F0 is the closest to the average of the network as a whole. The largest differences are those of the speakers with highest and lowest mean F0, Rich (29.5 Hz higher than the group average) and Pete (28.4 Hz lower than average).

Figures 1 and 2 display these differential data (on the x axis) against the recognition rate for each sample obtained in the SR test (on the y axis).

The close relationship between recognition rate and F0 scores become apparent on observation of these two figures. Those speakers with the highest differences from the group average are on the whole more successfully identified than those whose values settle closer to the group average. In short, speakers are more frequently identified when they have very high or very low F0, and when they have very high or very low standard deviations (indicating a wide pitch range or a monotonous voice respectively). These observations are shown to be statistically significant via a one-tailed Pearson product moment correlation test ($r = .692$, $df = 7$, $p < .025$ for mean F0; $r = .755$, $df = 7$, $p < .01$ for standard deviation).

DISCUSSION

Our analysis of F0 measures shows that both mean F0 and standard deviation of F0 have a statistically significant correlation with recognition rate in the SR test. It appears easier for listeners to identify those speakers who exhibit more extreme F0 levels or ranges. Those speakers who exhibit average F0 levels or ranges prove generally harder to identify.

We should not, of course, overstate the contribution of this F0 analysis, particularly with such a small corpus of speakers and listeners. It is abundantly clear that segmental factors play a major part in enabling listeners to recognize voices, as likewise do other suprasegmental factors (the presence of an innovative intonation pattern, and creaky phonation), at least in some cases. However, our data do at least provide preliminary evi-

dence that pitch patterns may make a positive contribution for the naïve listener approaching a SR task, just as they do for a forensic phonetician engaged in a comparison of speech samples (e.g. Hollien 1990: 196ff.; Baldwin and French 1990: 45; French 1994; Künzel 1997).⁴

Our data may therefore offer new information about the cognitive mechanisms of naïve SR, at least for telephone speech. It is commonly noted that speaker recognition takes place very quickly, within a second or so of the speaker being heard (e.g. Pollack *et al.* 1954; LaRiviere 1975; Baldwin and French 1990: 20). It has also been shown that increasing the length of samples has little effect on SR success rate (Pollack *et al.* 1954; Stevens *et al.* 1968). However, if we are right in our conclusion, it suggests that the SR task is not necessarily achieved with reference only to the first few seconds of a heard sample. F0 mean and range are, by definition, properties of an utterance or a whole spoken text. The fact that our listeners' responses do seem to be linked to F0 suggests that speakers may be able, at least in part, to make their judgements of speaker identity by processing phonetic information over a relatively long time domain. It remains to be investigated just how important such information may be in comparison with segmental information, and also how long a domain is relevant to listeners in processing the F0 information. It may also be the case that listeners presented with telephone speech are more sensitive to such information than those in direct contact, since it is to be expected that all possible cues will be analysed when speech is significantly degraded.

CONCLUSION

Our study has shown similar results to that of McClelland (2000). In a speaker recognition task using telephone speech, close friends are sometimes unable to identify each other and even themselves, and they make incorrect judgements of identity. Foil voices may be wrongly identified as in-group members. As previous studies have shown, the recognition task makes use of segmental and suprasegmental information. Highly salient features of certain voices make them more readily identifiable than others. We have also shown that F0 values, information pertaining to a long time domain, may also play a significant role in speaker recognition. Further research is needed to explore the influence of F0 in greater depth.

Finally, no conclusive evidence shows why listener performance in the SR test varies so much. Perhaps this is a skill which is affected by an individual's general sociolinguistic awareness and overt awareness of linguistic factors. It has also been suggested that individuals need time and experience to build up a repertoire of specific knowledge in order to recognize voices over the telephone, even when those voices are well known from direct interaction. Further research might profitably focus on this proposal, to assess the effect of different levels of experience on the task.

NOTES

- 1 Our thanks to Marion Shirt and Bethan Davies for their helpful comments on drafts of this article.
- 2 The technical effects of mobile phones may well differ from those of land-lines. Mobile phone networks may have the analogue signal converted into a digital one and then sent half way round the world before it is converted back into an analogue signal for transmission to a land-line. Several networks boast 'EFR' (Enhanced Full Rate), which is a technical means by which the signal is improved. This enhancement procedure no doubt introduces effects into the signal which differ from those produced by land-lines. Such differences have so far remained largely untouched by research in forensic linguistics, although McClelland (2000) found that the average F0 used by callers on mobile phones was as much as 30 Hz higher than that used by the same speakers using land-lines.
- 3 It is possible that stylistic shifts brought about by the phone context itself are responsible for these speakers' use of high rising tone, and likewise, for other speakers, creaky phonation. Although we do not have comparable data from non-telephone speech, such differences would offer further clues as to why close friends may be unable to recognize each other over the telephone. If these features are indeed less frequent in the subjects' regular speech style, their network colleagues may not have had enough experience of hearing the telephone-induced features in order to construct a sufficiently robust telephone recognition strategy (Baldwin and French 1990: 111). However, such possibilities obviously require further research.
- 4 In SR tests using very short samples (1.25 secs), LaRiviere (1975) showed that fundamental frequency was an equally good predictor of speaker identification and cross-speaker confusion as vowel formant frequencies. However, none of his results for these particular comparisons reached statistical significance. In light of our findings, it might be of interest to replicate a study such as LaRiviere's, using samples of longer duration.

REFERENCES

- Baldwin, J. and French, J. P. (1990) *Forensic Phonetics*, London: Pinter.
- Bradbury, M. (1996) 'It's goodbye Memsahib, hello Sheila', *Daily Mail*, 20 March, p. 8.
- Braun, A. (1995) 'Fundamental frequency – how speaker-specific is it?' in A. Braun and J. P. Köster (eds), *Studies in Forensic Phonetics*, Trier: Akademischer Verlag, 9–23.
- Cruttenden, A. (1997) *Intonation* (2nd edn), Cambridge: Cambridge University Press.
- Docherty, G. J. and Foulkes, P. (1999) 'Newcastle upon Tyne and Derby: instrumental phonetics and variationist studies', in P. Foulkes and G. J. Docherty (eds), *Urban Voices: Accent Studies in the British Isles*, London: Edward Arnold, 47–71.

- French, J. P. (1994) 'An overview of forensic phonetics with particular reference to speaker identification', *Forensic Linguistics*, 1: 169–81.
- Hirson, A., French, J. P. and Howard, D. (1994) 'Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics' in J. Windsor Lewis, (ed.), *Studies in General and English Phonetics: Essays in honour of J. D. O'Connor*, London/New York: Routledge, 230–40.
- Hollien, H. (1990) *The Acoustics of Crime: The new science of forensic phonetics*, New York/London: Plenum Press.
- Hollien, H. (1996) 'Consideration of guidelines for earwitness lineups', *Forensic Linguistics*, 3: 14–23.
- Künzel, H. J. (1990) *Phonetische Untersuchungen zur Sprecher-Erkennung durch Linguistisch Naïve Personen*, Stuttgart: Steiner.
- Künzel, H. J. (1994) 'On the problem of speaker identification by victims and witnesses', *Forensic Linguistics*, 1: 45–58.
- Künzel, H. J. (1995) 'Field procedures in forensic speaker recognition', in J. Windsor Lewis, (ed.), *Studies in General and English Phonetics*, London: Routledge, 68–84.
- Künzel, H. J. (1997) 'Some general phonetic and forensic aspects of speaking tempo', *Forensic Linguistics*, 4: 48–83.
- Ladefoged, P. (1993) *A Course in Phonetics* (3rd edn), Fort Worth: Harcourt Brace Jovanovich.
- LaRiviere, C. (1975) 'Contributions of fundamental frequency and formant frequencies to speaker identification', *Phonetica*, 31: 185–97.
- Laver, J. (1981) 'Linguistic routines and politeness in greeting and parting' in F. Coulmas (ed.), *Conversational Routines*, The Hague: Mouton.
- McClelland, E. (2000) 'Familial similarity in voices', paper presented at the BAAP Colloquium, University of Glasgow, April.
- Nolan, F. J. (1983) *The Phonetic Bases of Speaker Recognition*, Cambridge: Cambridge University Press.
- Nolan, F. J. (1991) 'Forensic phonetics', *Journal of Linguistics*, 27: 483–93.
- Nolan, F. J. (1997) 'Speaker recognition and forensic phonetics', in W. J. Hardcastle and J. Laver (eds), *The Handbook of Phonetic Sciences*, Oxford: Blackwell, 744–67.
- Nolan, F. and Grabe, E. (1996) 'Preparing a voice lineup', *Forensic Linguistics*, 3: 74–94.
- Pollack, I., Pickett, J. M. and Sumbly, W. H. (1954) 'On the identification of speakers by voice', *Journal of the Acoustical Society of America*, 26: 403–6.
- Shirt, M. (1984) 'An auditory speaker recognition experiment', *Proceedings of the Institute of Acoustics Conference*, 6: 101–4.
- Stevens, K. N., Williams, C. E., Carbonell, J. R. and Woods, B. (1968) 'Speaker authentication and identification: a comparison of spectro-

- graphic and auditory presentations of speech material', *Journal of the Acoustical Society of America*, 44: 1596–607.
- Tollfree, L. (1999) 'South East London English: discrete *versus* continuous modelling of consonantal reduction', in P. Foulkes and G. J. Docherty (eds), *Urban Voices*, pp. 163–84.
- Watt, D. J. L. and Milroy, L. (1999) 'Patterns of variation and change in three Tyneside vowels: is this dialect levelling?' in P. Foulkes and G. J. Docherty (eds), *Urban Voices*, pp. 25–46.
- Wells, J. C. (1982) *Accents of English* (3 Vols.), Cambridge: Cambridge University Press.