UDK 81'342.1
81'34:343.98
343.982.32:81
Pregledni rad

Francis Nolan
Department of Linguistics, University of Cambridge, UK

VOICE QUALITY AND FORENSIC SPEAKER IDENTIFICATION

SUMMARY

When phoneticians compare forensic speech samples they often remark in their reports on a "similarity of voice quality". Likewise, when earwitnesses are asked to describe a voice they have heard, they will normally comment on the accent, if they are able to, and additionally describe what they heard as an "X voice" where "X" is a term such as "rough" or "resonant" that can be seen as an informal label of voice quality. In this talk I will examine these two main categories of forensic speaker identification – by phonetic experts and by earwitnesses – with reference to the notion of voice quality. I will take voice quality in the broad sense discussed by Laver in his The Phonetic Description of Voice Quality *(CUP, 1980), that is, as covering supralaryngeal as well as laryngeal characteristics which emerge cumulatively from a person's speech.*

In speaker comparison by phonetic experts the emphasis in acoustic analysis tends to be on segmental properties, or on pitch-related long-term features. I will give some examples of how speakers can be differentiated in this way, and touch on how the dynamics of formants in transitional parts of the speech signal may provide the nearest we have to a speaker's "signature". Beyond segmental analysis, however, I will show that an analysis using the long-term distributions of formant frequencies can capture information relating to Laver's supralaryngeal voice quality categories. Given the availability of Laver's comprehensive framework for the impressionistic analysis of voice quality we might ask why, in the auditory strand of their forensic analyses, phoneticians have made little use of systematic voice quality description, and I will explain why I think that is.

As regards earwitness evidence I will focus on the description of voices by earwitnesses, and on the use of voice parades. I will ask whether an earwitness's description of a voice might be improved if questioning of the witness were informed and structured by knowledge of a framework for voice *quality description. And in creating a voice parade, I will show how pre-tests are used to ensure that the parade is fair, including one where experimental subjects are, in effect, asked to rate the similarity in voice quality between all pairs of samples to be used in the parade. This is to ensure that the suspect is not an outlier. Finally I will preview a project which will investigate the effect of the telephone on such similarity judgments.*

Key words: *voice quality, speaker identification, forensic phonetics*

THE MEANINGS AND DESCRIPTION OF "VOICE QUALITY"

The term "voice quality" is used in several different ways by phoneticians. It can refer narrowly to the effect resulting from the mode of vibration of a person's vocal cords, or to the total perceptual effect of a speaker's vocal activity. It can mean an auditory effect which persists through all a speaker's vocal output, or to episodes where a temporary modulation of that output takes place, as when a person uses breathy voice to indicate a confidential item, or a palatalised, lip-rounded setting in baby-talk. And, thirdly, we need to distinguish conceptually between voice quality as something which is under a speaker's control (as of course the adult's "baby-talk" setting is), and voice quality as an effect determined by a speaker's physiology and therefore beyond his or her control.

Laver (1980) provides probably the most comprehensive linguistic-phonetic framework for the description of voice quality. He takes voice quality to cover both laryngeal and supralaryngeal effects, and indeed effects too which arise from overall settings of the vocal organs, such as their degree of tension. He allows voice quality to include both the background quality which pervades all a person's speech, and short or medium term modulations of that background by the temporary adoption of different settings of the vocal organs. And the framework he devises is one which explicitly describes all the effects that anyone with a normal vocal apparatus can (in principle) produce, much as the IPA framework describes those segmental effects which are in principle achievable by any vocal tract. Laver's voice quality framework is first and foremost a tool in the toolbox of linguistic-phonetic description, allowing us to describe, say, how one dialect may differ from another by being relatively denasalised, how one sociolect may be characterised as using harsh whispery phonation where another tends towards creaky voice, or how a given language might use palatalisation as a paralinguistic resource to express politeness.

Given the orientation of Laver's framework towards the description of linguistic and paralinguistic effects which all physiologically normal vocal tracts are able to achieve, one might assume *a priori* that it has no role in the description of voice quality as it marks an individual speaker. On the contrary, the framework can in fact be used to describe individual voice quality, and it has been so used. One reason for this is that a speaker's "characteristic, quasi-permanent, auditory colouring" (as Laver has termed it) is not just a product of that person's vocal anatomy, but of how he or she habitually uses it. Two speakers might have (let us suppose) anatomically identical vocal tracts – perhaps this really does happen in the case of identical twins – and speak exactly the same dialect, yet differ in how they sound because one has a tendency to speak with a tense laryngeal setting and slightly raised larynx position, and the other with a lax, breathy laryngeal setting and a lowered larynx position.

Another reason for the applicability of the voice quality framework to individual voice quality is that the clear conceptual distinction between what arises from anatomy and what arises from vocal behaviour is far from clearly observable in speech data. On hearing a speaker who is markedly denasal compared to the rest of the speech community we cannot be certain whether this arises from a permanent obstruction of the velic opening by adenoids, from a temporary obstruction as a result of an infection, or from an idiosyncratic learned speaking habit. The terms in the descriptive framework describe effects which are, in principle, under the speaker's control, but which can also arise from (or be mimicked by) anatomical characteristics. Perhaps the most dramatic example of this kind of crossover into the description of characteristics arising from anatomy is the work of Beck (1988, 1997), who used Laver's voice quality framework to describe the speech of Down's Syndrome speakers. She showed that Down's Syndrome speech was associated with high degrees of an auditory effect attributed in the framework to a tongue-body setting of palatalisation. This auditory effect turns out to reflect an anatomical tendency to an underdeveloped palatal arch relative to the size of the tongue. It is obviously not correct to infer from the description of the speech as palatalised that the Down's Syndrome speakers are 'choosing' to palatalise their speech, but nonetheless the framework allows, as does the description of a vowel in terms of the vowel quadrilateral, an auditory impression to be consistently captured and conveyed analytically.

Given that Laver's framework, or indeed any linguistic-phonetic framework for describing voice quality that might be devised, can be applied to the sound of a voice without the effects being attributed to an origin in volitional behaviour versus anatomy, it would seem reasonable to expect that it would have wide application in forensic phonetics.

SPEAKER IDENTITY IN FORENSIC PHONETICS

A substantial part of the activity which goes on under the heading of "forensic phonetics" has to do with the relation between samples of the voice and the identity of the speaker. On the one hand, there is the task which has traditionally been called "forensic speaker identification" in which a phonetician is (typically) asked to compare a reference sample of a suspect's speech with a sample recorded in connection with a crime – for instance a bomb warning, a telephoned threat, or a fraudulent telephone transaction with a bank.

The party commissioning the comparison, whether the investigating authority (often the police), the prosecuting authority, or the defence, would always prefer a nice simple answer in the form of an identification or elimination, hence the traditional term "speaker *identification*". However I and others have repeatedly stressed in publications the variable relation between an individual and the acoustic speech signal, this variability resulting from the plasticity of both the vocal mechanism (the speech organs) and the linguistic system (which allows for style shifting, dialect modification, and so on). In the UK, therefore, forensic phoneticians have come to prefer the term "speaker *comparison*", and have agreed a way of formulating conclusions which reflects the limitations of forensic identification by voice.

Very briefly, instead of using a five point likelihood scale for identification or elimination (e. g. "5 almost certainly the same", "4 very like the same" etc.) the conclusion is split into two parts. First, a decision is made on whether the samples are consistent with having been spoken by the same person. There may be inconsistencies which cannot be explained by known models of variation (acoustic, articulatory, stylistic, sociolinguistic, etc.), in which case the samples are not "consistent", and that is all that is to be said. In the absence of such inexplicable inconsistencies, the samples are "consistent", but that says no more than that it is *possible* that the two samples came from the same speaker. Second, the consistent features which the samples exhibit are assessed with respect to a five point scale of distinctiveness from "1 not distinctive" to "5 exceptionally distinctive". If the samples are short, all the acoustic values near the middle of their ranges in the population, and the accent is a perfectly unremarkable examplar of a widely spoken accent, then the consistent features are "not distinctive", and the evidence will be have no more weight than a failure to eliminate the individual, which is what it is. At the other extreme, if the samples are long and rich in unusual features, such as very low pitch and formants within the range of the population, an idiosyncratic stutter, another speech abnormality such as a lisp, and a highly unusual mixed accent (Hebridean Scottish mixed with Jamaican, say), the consistent features are "exceptionally distinctive" and the conclusion would in effect be a positive identification. But even there, the formulation of the conclusion forces the court to appreciate the complexity of the process, and no outcome can be mapped simplistically into a "guilty" verdict.[1]

THE FEATURES THAT ARE ANALYSED

In carrying out a speaker comparison, forensic phoneticians nowadays normally use both traditional auditory techniques and computer-based acoustic analysis. These provide complementary information: the ear (together with the brain!) is the best tool we have for carrying out linguistic-phonetic analysis of, in particular, dialect, whereas quantitative acoustic analysis can reveal speaker-distinguishing cues which the ear, usefully for most purposes, ignores (Nolan, 1993).

Some attention is paid to long term characteristics of the speech, such as fundamental frequency statistics (although these tend to be very sensitive to speaking style and environment – for instance background noise causes speakers to raise their fundamental frequency). Much of what is done is at the segmental level, whether auditory comparison of pronunciation, or acoustic analysis of properties such as formants. I personally believe that formant analysis is of great

[1] Or a "not guilty" verdict in the much rarer circumstance that a defendant being demonstrated to be the speaker on a recording exonerates him or her from being involved in a crime.

importance because formant values reflect the interaction of three potentially indentifying sources: the linguistic accent, the anatomy of the individual's vocal tract, and the speaker's acquired articulatory strategies.

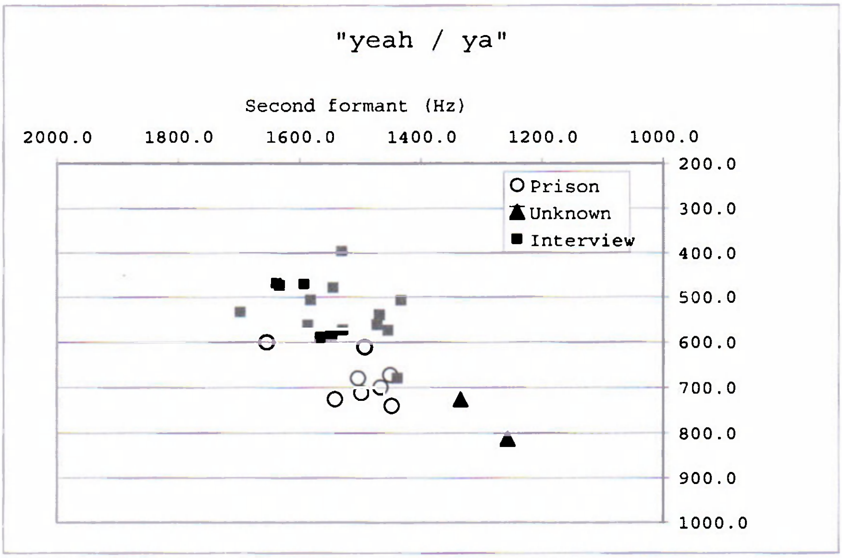

Figure 1. Two-formant plot of the vowel in "yeah/ya" showing two tokens from a telephoned bomb warning and numerous tokens from a suspect in interview and recorded covertly over the telephone from prison

Slika 1. Prikaz vrijednosti dvaju formanata vokala u primjerima "yeah/ya", od kojih su dva sa snimke telefonskog upozorenja o bombi, dok su ostali ili izgovoreni tijekom intervjua ili potajno snimljeni tijekom telefonskih razgovora iz zatvora

This is not the place to go into details, but as an example Figure 1 shows how, on a standard two-formant plot, two tokens of "yeah" from an unknown telephoned bomb warning (triangles) fall outside the range of a large number of tokens of the same word from a suspect. The suspect is represented by a substantial number of tokens recorded direct in his police interview, and eight tokens from covertly recorded telephone calls made by the suspect from prison. The two larger clusters illustrate one of the complications of using formant frequencies, namely that values are affected by the bandwidth limitation of the telephone (Künzel, 2001; Rose, 2003), so that the clusters from the two samples known to be from the suspect are rather different in terms of their first formant. However Künzel shows that F2 values are generally well represented by the telephone, and this plot shows clearly that F2 in the bomb call is lower than in any of suspect's utterances, as well as F1 tending to be higher than in the other (prison) telephone calls. A single observation of this kind is only one piece in a jigsaw puzzle, but this kind of inconsistency recurring over a substantial number of vowels or other sounds examined weighs heavily in the direction of concluding that the samples are not consistent with having been spoken by the same speaker.

What this kind of analysis does not capture, based as it is on formants measured at one instant in a vowel, is the dynamically changing configuration of formants through time. Given that formant frequencies at any moment result from the interplay of the three factors mentioned above – accent, anatomy, and acquired articulatory strategies – it is unlikely that any two individuals, even if they coincide at one point in a sound, will shadow each other through time. One of the hypotheses being explored currently in the DyViS project in Cambridge, and elsewhere in the work of Kirsty McDougall, is that it is in the dynamics of formant trajectories that a speaker's "vocal signature" lies.

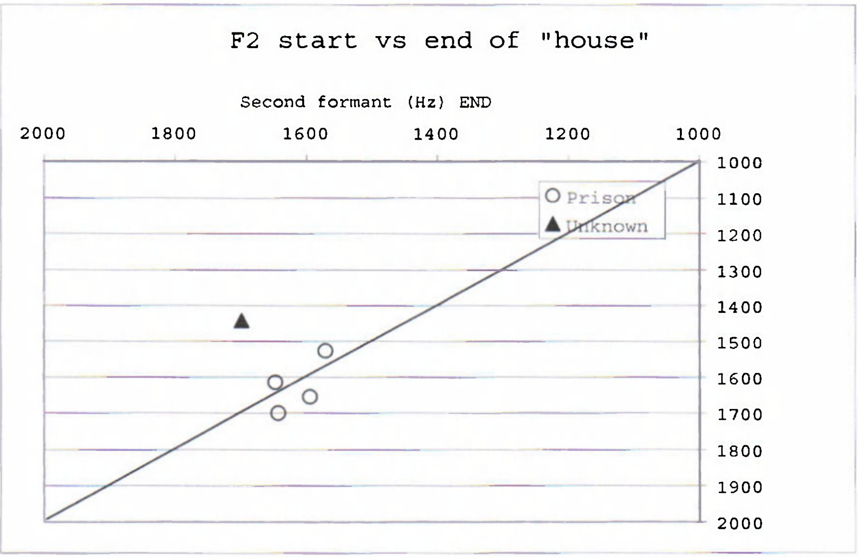

Figure 2. Plot of F2 at the start and end of the vowel in "house" (*not* F1 against F2) from a telephoned bomb warning (triangle) and four tokens from a suspect recorded over the telephone from prison

Slika 2. Prikaz vrijednosti F2 na početku i na kraju vokala u riječi "kuća" (*ne* odnos F1 i F2) iz telefonskog upozorenja o bombi (trokut) i četiriju primjera sa snimki telefonskih razgovora osumnjičenog iz zatvora

So far, in practical forensic work, attempts to capture this dynamism have been unsophisticated. Figure 2 plots F2 at the start and end of the vowel/diphthong in the word "house" (*not* F1 vs. F2), recorded over the telephone, for the same bomb warning and suspect's prison calls as in Figure 1. The data are very limited, but reveal that the one available token from the bomb caller is diphthongal, and separate from the four tokens from the suspect, which lie on the diagonal and are therefore monophthongal. Again, by itself, this graph is merely another piece in the jigsaw puzzle, but it contributes to an overall picture where the bomb caller too often lies outside the range of the suspect for the recordings to be considered to be consistent with having been spoken by the same speaker – even though the quantisation of the dynamic evolution of formants is crude.

In future, however, we can expect the kinds of technique being developed by Kirsty McDougall to find their way into forensic casework. Here, the whole formant trajectory in a dynamically changing sequence (such as a diphthong followed by a particular consonant, or a vowel-liquid-vowel sequence) is tracked, and, in the most recent implementation (McDougall & Nolan, 2007) the tracks are modelled with a polynomial equation to give a compact descriptor of the trajectories. The trajectories from different speakers are then compared using discriminant analysis. The technique proves to have considerable potential for discriminating speakers. Research is proceeding on this and other aspects of speaker comparison in the context of our DyViS project.[2]

VOICE QUALITY DESCRIPTION IN SPEAKER COMPARISON BY EXPERTS

Valuable as such analysis is, one might be tempted to ask "isn't this missing the point? Shouldn't we be looking for descriptors of voice quality, which, after all, is what people are getting at when they describe a voice as 'resonant' or 'deep' or 'pleasant'?" These are fair questions; and, particularly given the existence of an extremely sophisticated framework for the description of voice quality, we need to consider how far this framework has been applied in forensic speaker comparison – and why it has not been more widely employed.

In Nolan (2005) I cited a revealing statistic: of 30 forensic speaker comparison cases that I had been involved in up to the time of writing, only two included an explicit reference to terms within a voice quality framework. A third – one of my own – included a reference to falsetto phonation, but more as an intermittent paralinguistic effect. Many reports did include a rather perfunctory reference to "a similarity in voice quality" but they did not elaborate on what analysis, other than a general impression, led to this statement, or what the claimed similarity consisted in.

[2] "Dynamic variability in speech: a forensic phonetic study of British English" funded by the UK Economic and Social Research Council, Grant RES-000-23-1248. For more details, see: http://www.ling.cam.ac.uk/dyvis/

In discussing the superficially surprising lack of systematic voice quality analysis in forensic speaker comparison I considered, and rejected, the possibility that experts were not aware of the availability of descriptive frameworks for voice quality. This is unlikely, and certainly cannot be the reason that I myself have very rarely applied Laver's framework since an important part of my early research on speaker characteristics (Nolan, 1983) involved measuring the acoustic correlates of different voice quality components. More likely, I suggested, was a lack of training. Whilst the IPA framework for segments is a standard element in phonetic training, Laver's (or any other) voice quality framework is generally not; and yet to use it successfully a phonetician needs to have the same kind of guided auditory and productive training in order to internalise the analytic perceptual categories and associate them with their articulatory correlates as are needed in the case of the Cardinal Vowels. Laver and his colleagues have run quite a number of training courses, but these have been principally attended by speech pathologists rather than forensic phoneticians. Thirdly, even if an expert has the necessary training, applying the framework is a time-consuming business and might not be given priority over more quantitative analyses.

However the main reason for the lack of take up of linguistic-phonetic voice quality description, I proposed, was the limitations imposed by the telephone. In the majority of speaker comparison cases the disputed sample is recorded over the telephone, and of course the telephone limits and potentially distorts the speech signal. Sound energy below about 300 Hz, and above about 3 500 Hz, is lost, and there may be distortions of the spectral shape particularly in the vicinity of these cut-offs. For reference, the first harmonic of a male voice may be as low as 75 Hz, and significant fricative energy may be present up to 10 000 Hz. Calls which are routinely recorded, such as those to the emergency services or to banks, are recorded on bulk recorders which may further degrade the signal, as may answering machines or hand-held recorders used by people trying to record telephone messages. All in all, the sample of speech from the "unknown" is like to be of very poor quality compared to anything a phonetician or linguist would normally encounter in research.

It's worth reminding ourselves here that this degradation of the speech signal limits most kinds of forensic phonetic analysis. There is no possibility of doing a phonetic comparison of all the speech sounds of a language, since sounds such as fricatives and stop bursts, for instance, will have lost most of their (high frequency) energy, and the ear cannot restore what is not there.

As far as voice quality is concerned, we need to consider separately the laryngeal and supralaryngeal contributions. A speaker's laryngeal setting determines the voicing source spectrum. For one thing, the breathier the phonation, the sharper the fall-off in energy in successively higher harmonics, and (up to a point) the tenser the phonation the shallower the fall-off. Nolan (1983:142-55) shows the effect on the long term average spectrum of adopting various of the laryngeal components in Laver's system. Another common quantification (Ní Chasaide & Gobl, 1997:442-43) is the ratio of the first harmonic to the second harmonic or to other higher harmonics; the breathier the voice, the greater the dominance of the first harmonic. For another thing, some laryngeal settings, especially breathy and whispery settings, generate high frequency aperiodic energy ("noise") in parallel with the voice source. All of these acoustic manifestations of laryngeal voice quality will be distorted by the bandwidth limitation inherent in telephone and telephone-like transmission.

If judgments are to be made about the laryngeal voice quality of such a sample, they can only be made via a rather elaborate process of perceptually reconstructing what the sample would have sounded like had it not been passed through the telephone. Undoubtedly we have quite a bit of skill in doing this, since we are generally able to associate the voice of a familiar person over the telephone with that person, and indeed we do, with varying degrees of accuracy, recognise callers; but it remains to be demonstrated that the componential and independent judgments involved in the auditory analysis of laryngeal voice quality could be accurately carried out in a way which compensates for the effects of bandwidth limitation.

Are supralaryngeal settings, whose effects are found mainly in the relative frequencies of formants (Nolan, 1983), a more reliably perceivable element of voice quality in the forensic situation? The answer is probably "yes", but even here we must expect some problems. As noted in Section 3 above, formants which are near the limits of the 300-3 500 telephone bandwidth will not be accurately represented, and this effect could in principle result in changes in the perception of supralaryngeal settings. For instance, just as Künzel (2001:94) suggests that /iː/ will sound like /ɪ/ over the telephone, because its F1 frequency is raised by the loss of low frequency energy, so we might fear that a strongly palatalised supralaryngeal setting might be less noticeable[3]. Once again, whether the telephone signal retains enough information for the true voice quality setting of the speech to be reconstructed is an empirical matter on which research is needed.

Until appropriate experiments have been carried out, it seems we need to be cautious about the scope for accurate analysis of voice quality components (such as breathiness, palatalisation, and so on) with band-limited speech. Equally, if the information is not there to do componential analysis, we need to treat with a degree of scepticism global judgments often encountered in forensic reports such as "a good match in voice quality between the telephone call and the suspect's sample". The attempt to make such a judgment may often involve a

[3] Interestingly, a recent experiment in Cambridge found that phonetically trained listeners plotting vowels from direct and telephone recordings actually seemed to compensate for this telephone effect, contrary to Künzel's prediction, though explanations other than compensation have yet to be excluded. (Lawrence, S., Nolan, F., & McDougall, K. "Acoustic and perceptual effects of telephone transmission on vowel quality". To be submitted to *International Journal of Speech, Language and the Law*.)

dangerous leap of faith. The situation is perhaps not much better than in the case of fricative comparison involving telephone samples.

A QUANTITATIVE TOOL FOR SUPRALARYNGEAL VOICE QUALITY

Fairly recently, Catalin Grigoras has developed a neat method for capturing long term resonance properties in speech (e. g. Nolan & Grigoras, 2005). This uses Linear Prediction analysis to estimate the formant frequencies at each relevant time-frame through the course of a speech sample, and then the statistical distribution of the formant frequencies is plotted. Figure 3 shows an example of such an analysis.

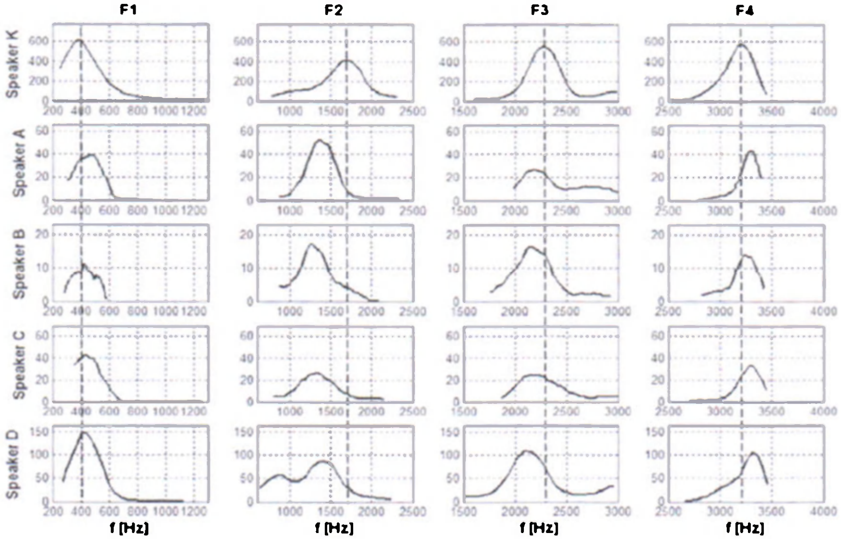

Figure 3. Output of long term formant analysis (taken from Nolan & Grigoras, 2005) showing statistical distributions of formant estimates. Speaker K is a suspect speaking over the telephone, and "speakers" A-D are samples from four obscene telephone calls, likely to be made by the same individual. The vertical dotted lines facilitate comparison of the centre of the suspect's formant distributions with those from the unknown speaker(s).

Slika 3. Rezultati dugotrajne formantske analize (prema Nolan i Grigoras, 2005) pokazuju statističku distribuciju očekivanih vrijednosti formanata. Govornik K je osumnjičenik koji razgovara preko telefona, a "govornici" od A do D primjeri su iz četiriju obscenih telefonskih razgovora koje je vjerojatno počinio isti osumnjičenik. Okomite iscrtkane linije olakšavaju usporedbu središta formantskih distribucija osumnjičenika s onima nepoznatog/-ih govornika.

Each pane in Figure 3 shows a statistical distribution of the formant frequency estimates for a particular formant (F1 to F4) computed over many seconds of speech. "Speaker K" is the reference sample of a suspect. Fortuitously, since the unknown samples are telephone speech, this reference sample is also a (covert) telephone recording. "Speakers A-D" are samples from four obscene telephone calls, likely to have been made by the same individual. It can readily be seen that the suspect's second and third formants are considerably higher than those of the obscene calls. This difference in long term resonance properties, together with a number of non-matching segmental formant frequencies, make it possible to say that the suspect's speech is not consistent with the speech recorded from the obscene telephone calls.

Most interestingly for the present discussion of voice quality, the suspect's higher F2 and F3 suggest a greater degree of palatalisation, and this is consistent with the overall perceptual effect of his voice quality. In this case we might expect Speaker K's F1 to be lower, but again we have to remember the "telephone effect" which will tend to inhibit low F1 estimates even where the real F1 is low. The telephone effect, of course, means that even more careful interpretation would be required in situations where a telephone sample is being compared with a directly recorded sample.

Despite the caution needed in the context of telephone speech, it seems to me that Grigoras's long term formant method deserves serious testing and development both as a means of advancing the quantitative characterisation of the descriptive terms for voice quality settings begun in Nolan (1983), and as a practical tool in forensic speaker comparison.

VOICE QUALITY, SPEAKER SIMILARITY, AND VOICE PARADES

Voice parades, or voice line-ups, are occasionally used to test whether a suspect is the person an earwitness heard in connection with a crime. The fact that, in a parade, the witness is presented with samples from a number of "foils" (usually about seven) as well as from the suspect provides a measure of protection to an innocent suspect: a witness who is trying to be helpful but in reality is guessing is more likely to pick a foil than the suspect. However that is only true if the parade – and this applies equally to the more traditional visual parade – is "fair", in the sense that the foils are appropriate to the description of the perpetrator of the crime, and the suspect does not in any way "stick out" from the group of foils, which might cause a guessing witness to focus on the suspect.

In practice, because earwitnesses' descriptions of voices tend to be rather sketchy, the emphasis tends to be on making sure that the foils are a fair match to the suspect; that is, avoiding the auditory equivalent of having a set of white European foils standing in a line-up with a black suspect.

Suspect and foil samples are normally directly recorded rather than either of them being telephone speech. In the UK, the practice has evolved of using a police interview of the suspect as a source of speech samples for the parade (the samples of course have to be scrupulously chosen to avoid any connection with the crime in terms of content); and, in order that speaking style is best matched, the foil samples are chosen from police interviews with other similar-sounding real suspects in unrelated cases.

Perhaps because full bandwidth recordings are available (though not necessarily of particularly good quality), it is in the preparation of voice parades that I have made some use of voice quality analysis forensically. In connection with the voice parade reported in Nolan and Grabe (1996:78-85) my working notes on the potential foil samples included comments such as "heavy nasalization", "tenser voice than [the suspect]", and "light voice" (not a term within Laver's framework, but one which I would have been using as a cover term for a voice with relatively high pitch and formant frequencies, and possibly phonation tending towards breathy rather than tenser settings). I also rated the distance in terms of voice quality (and separately accent) between each foil and the suspect on a numerical scale. Although this was not a thorough application of the framework, I at least had in mind the dimensions of voice quality systematized in Laver's (1980) framework.

Really, though, what needs to be assessed to ensure the fairness of a voice parade is not "voice quality", but "speaker similarity" – of which voice quality may be only one component. In many respects, the best arbiter of the fairness of a voice parade would be naïve listeners intuitively rating speaker similarity. Rietveld and Broeders (1991) demonstrate an experimental method for this, in which subjects rate all possible pairings of speakers (including same--same pairings, obviously using different samples) on a scale between (for instance) "very different" and "very similar". Multidimensional scaling is then used to reduce the data so that the speakers could be represented in a two--dimensional space, allowing the distances between them to be visualised.

This kind of speaker-similarity assessment has more recently been incorporated by Kirsty McDougall into the evolving procedure for voice parades. It provides a test of the "fairness" of the voice parade, and in particular checks that the suspect is not an "outlier". Fig. 4 shows the outcome of McDougall's test on her voice parade, which shows that her suspect lies comfortably in the same space as her foils.

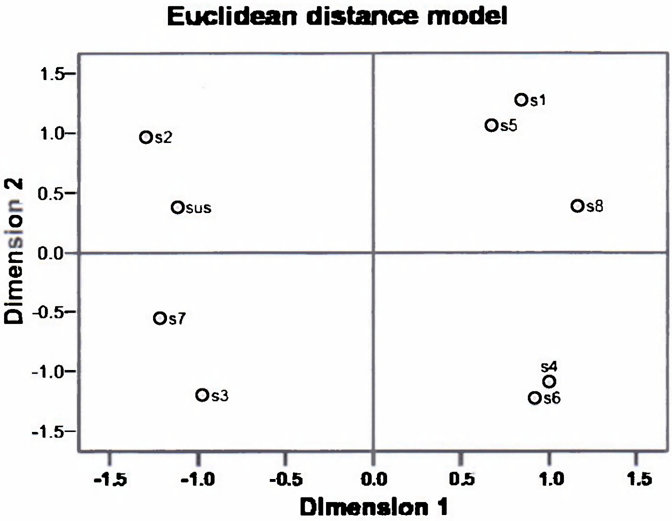

Figure 4. Two-dimensional transformation of distances established by rating all possible pairings of speakers in a candidate voice parade

Slika 4. Dvodimenzionalna transformacija udaljenosti dobivena ocjenjivanjem svih mogućih parova govornika u nizu glasova kandidata

We cannot disentangle the effects of voice quality from those of accent in such results, except that in all voice parades the phonetician choosing the samples will have exercised considerable care to ensure that the speakers are very similar in accent, and so we can presume that much of the distribution of the speakers in Fig. 4 arises from their personal voice quality (arising, of course, both from anatomical and volitional sources). As a small piece of evidence, a pilot experiment on the samples used in the parade in Nolan and Grabe (1996) showed a good agreement between my voice quality distance ratings and naïve subjects' dissimilarity ratings. However the relation between voice quality and speaker similarity is one which will require further research if we are to understand it more fully.

Finally, we need to return to the question of the effect of the telephone. It is not uncommon in everyday life to hear comments to the effect that the telephone affects the sound of a person's voice. Such comments may be along the lines of "oh, you sound different on the phone", or, not infrequently, that a person's accent sounds more noticeable than face-to-face. Given what we know about the modifications imposed on the speech signal by the telephone, it is *a priori* reasonable to predict some telephone effects on speaker similarity. It is not implausible that an earwitness may be asked to make an identification "cross--modally" – perhaps being previously familiar with an individual from speaking face-to-face, and then hearing a voice which might be that speaker making an incriminating phone-call; or listening to a voice parade of direct recordings in order to pick out a voice heard at the time of the crime over the telephone. Additionally, in a case where the identification involves exclusively telephone speech, it would be important to know whether speakers sound more similar over the phone (and therefore the risk of mistaken identification is higher).

We are about to investigate the effect of the telephone on voice similarity in a companion project[4] to the DyViS project (see footnote 2), which has collected a database of 100 aged-matched speakers of Standard Southern British English doing different speaking tasks. One of these tasks was a spontaneous conversation which was recorded direct and at the remote end of a telephone link. Fifteen similar-sounding speakers will be chosen from the DyViS database, initially on the basis of auditory judgements. A similarity rating experiment will be run, as described above. Naïve listeners will rate the perceived distances between all pairings of the 15 test speakers, including same-same speaker pairs. Three groups of listeners will hear the same linguistic material representing the speakers; one group will hear only directly recorded pairs of samples, another group will hear only telephone pairs, and a third group will hear only cross--modal pairs (direct-telephone or telephone-direct, randomly ordered). The results should tell us whether speakers are inherently more similar over the telephone, and whether cross-modal listening affects similarity judgments. Acoustic analysis will then allow us to estimate which parameters of the voice have most weight in similarity judgments, and whether some are more robust than others over the telephone.

CONCLUSION

This paper has explored a number of aspects of the relation between voice quality, considered as a global perceptual effect arising a speaker's speech, and the speaker-defining characteristics which are of central interest (albeit elusive) in forensic speaker comparison and earwitness identification. I have shown that a slightly uneasy association exists between the linguistic-phonetic analysis of voice quality and forensic practice, which tends to have focused more on segmental features. Systematic voice quality analysis hasn't really established itself in forensic practice, except perhaps marginally when phoneticians are selecting foil samples for voice parades, nor do forensic phoneticians have a clear model of how voice quality relates to ordinary hearers' experiences of how similar or dissimilar speakers are to each other.

[4] "Voice similarity and the effect of the telephone: a study of the implications for earwitness evidence" funded by UK Economic and Social Research Council Grant RES-000-22-2582, starting January 2008.

Nonetheless I hope I have shown that voice quality analysis has an important role to play in our understanding of speaker characteristics. Just as a forensic phonetician measuring vowel formant frequencies will integrate them with auditory impressions by means of the linguistic-phonetic model of vowel quality provided by the IPA, so we may come to interpret long term formant distributions (as in Fig. 3) and our auditory impressions of a speaker with respect to Laver's voice quality framework.

When it comes to naïve listeners, we know surprisingly little about what underlies their perception of speakers as similar or dissimilar. Experiments we plan to do in the near future using carefully matched speakers should tell us a lot about what acoustic dimensions listeners rely on to discriminate speakers, and whether these dimensions are affected by telephone transmission. The patterns which emerge in the results of these experiments would provide an illuminating comparison with an auditory voice quality analysis of the speakers used. This is not part of the planned project, but might be a profitable route to pursue towards the goal of reconciling voice quality analysis and forensic speaker characterisation.

REFERENCES

Beck, J. M. (1988). Organic variation and voice quality. PhD dissertation, University of Edinburgh.

Beck, J. M. (1997). Organic variation of the vocal apparatus. In W. J. Hardcastle and J. Laver (eds.), *A Handbook of Phonetic Sciences*, 256-297. Oxford: Blackwell.

Künzel, H. J. (2001). Beware of the "telephone effect": the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics* 8(1), 80-99.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: CUP.

McDougall, K. and Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. *Proceedings of the 16th International Congress of Phonetic Sciences,* 1825-1828. Saarbrücken: Universitat des Saarlandes.

Ní Chasaide, A. and Gobl, C. (1997). Voice source variation. In W. J. Hardcastle and J. Laver (eds.), *A Handbook of Phonetic Sciences*, 427--461. Oxford: Blackwell.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: CUP.

Nolan, F. (1993). Auditory and acoustic analysis in speaker recognition. In J. Gibbons (ed.), *Language and the Law*. London: Longman.

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle & J. Beck (eds.), *A Figure of Speech: a Festschrift for John Laver*, 385-411. Mahwah, New Jersey: Erlbaum.

Nolan, F. and Grabe, E. (1996). Preparing a voice line-up. *Forensic Linguistics* 3(1), 74-94.

Nolan, F. and Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law* 12(2), 143-173.

Rietveld, A. C. M. and Broeders, A. P. A. (1991). Testing the fairness of voice identity parades: the similarity criterion. *Proceedings of the 12th International Congress of Phonetic Sciences*, 5, 46-49. Universite de Provence, Aix-en-Provcnce.

Rose, P. (2003). The technical comparison of forensic voice samples. In I. Frcckleton and H. Selby (eds.), *Expert Evidence*, Chapter 99. Sydney: Lawbook Co.

Francis Nolan
Odsjek za lingvistiku, Sveučilište u Cambridgeu, UK

GLASOVA KVALITETA I FORENZIČKO PREPOZNAVANJE GOVORNIKA

SAŽETAK

Kad fonetičari uspoređuju forenzičke govorne uzorke, često u svojim izvještajima govore o "sličnosti glasove kvalitete". Isto tako, kada se od svjedoka traži da opišu glasove koje su čuli, oni, ako mogu, obično komentiraju akcent te dodatno opisuju ono što su čuli kao "X glas", pri čemu je X termin poput "grub" ili "rezonantan" i koji se može shvatiti kao neformalna oznaka kvalitete glasa. U ovome ću predavanju razložiti te dvije glavne kategorije forenzičkog prepoznavanja govornika – uz pomoć fonetskih stručnjaka i uz pomoć svjedoka – oslanjajući se na pojam kvalitete glasa. Kvalitetu glasa shvatit ću u njezinu širokom smislu predstavljenom u The Phonetic Description of Voice Quality *(CUP, 1980) Johna Lavera, odnosno kao kategoriju koja obuhvaća kako laringalne tako i supralaringalne osobine koje su zajednički rezultat govora neke osobe.*

Pri usporedbi govornika fonetski stručnjaci u akustičkoj analizi naglasak stavljaju na segmentalne osobine ili na dugotrajne osobine povezane s osnovnim tonom. Ilustrirat ću kako se govornici mogu razlikovati na ovaj način te dodirnuti temu dinamike formanata u prijelaznim dijelovima govornog signala, koji su možda najbliži ekvivalent govornikovu "potpisu". Osim segmentalne analize, pokazat ću da analiza dugotrajnih distribucija formantnih frekvencija može pokazati informacije povezane s Laverovim supralaringalnim kategorijama glasove kvalitete. S obzirom na dostupnost Laverova sveobuhvatnog protokola za impresionističku analizu glasove kvalitete, nameće se pitanje zbog čega, u auditornom dijelu forenzičke analize, fonetičari slabo iskorištavaju sustavni opis glasove kvalitete. Pokušat ću ponuditi odgovor na to pitanje.

Što se tiče dokaza svjedoka, koncentrirat ću se na svjedokove opise glasova i na uporabu nizova glasova. Postavit ću pitanje može li se svjedokov opis glasa poboljšati ispitivanjem koje je informirano i strukturirano poznavanjem protokola za opis glasove kvalitete. Pri stvaranju nizova glasova pokazat ću na koji način koristiti predtestove da bismo osigurali pravednost, uključujući one u kojima se eksperimentalne ispitanike traži da procijene sličnost kvalitete glasa svih parova u uzorku koji se koristi u nizu. Ovim se postupkom provjerava odstupa li osumnjičenik od prosjeka. Konačno, predstavit ću projekt kojim će se istraživati utjecaj telefona na takve sudove o sličnosti.

Ključne riječi: *kvaliteta glasa, identifikacija govornika, forenzička fonetika*