Paper number 1

Speaker-Specific Patterns of Frication of Voiceless Stops in Australian English

Deborah Loakes and Kirsty McDougall

Frication of stops occurs when a phonemic stop is produced with an incomplete closure, such that the stop is effectively reduced to a fricative (e.g. Shockey 2003). This is a typical connected speech process in dialects of British English (e.g. Knowles 1973, Nolan and Kerswill 1991, Shockey 2003) and has also be observed in Australian English (e.g. Tollfree 2001).

The present study examines speaker variability in patterns of frication of the voiceless stops /k/ and /p/ in Australian English. The study was prompted by initial auditory analyses of spontaneous speech data in which fricative realisations of /k/ and /p/ appeared to occur to varying extents for different speakers. Although speaker-related variability in the distribution of different realisations of voiceless stops has been noted in some studies, e.g. Sangster (2002) regarding Liverpool English, its potential use as a conveyor of speaker-characterising information has not been explored in detail.

The study is part of a broader investigation of speaker variability in the speech of twins, for which four pairs of Australian English-speaking male twins have been recorded. The twins each participated separately in two non-contemporaneous recording sessions, engaging in spontaneous conversation with the first author. Acoustic waveforms and wideband spectrograms of all /k/ and /p/ tokens produced by the speakers were inspected. Each token was labelled as closed or fricated, and the proportion of fricated tokens calculated for each consonant, separately for each speaker's two recording sessions.

For both /k/ and /p/, the proportion of fricated tokens was relatively consistent for a given speaker across the two recording sessions. The proportions of fricated /k/ and fricated /p/ tokens varies extensively, both among different speakers in general, and within the twin pairs.

The implications of these findings for forensic speaker identification will be discussed, as well as their relevance to theories of connected speech processes and sociophonetics.

Deborah Loakes is with the University of Melbourne, Australia.
Kirsty McDougall is with the University of Cambridge, United Kingdom.

Paper accepted (yes/no):

Paper number 2

Results of the 2003 NFI/TNO Forensic Speaker Recognition Evaluation

J.S. Bouten

In the Netherlands police officers listening to or tapping into telephone conversations are often confronted with speakers speaking a non-Dutch language. One of the biggest problems this brings about is that for some police officers distinguishing between these speakers is next to impossible. In this paper the results of the NFI-TNO Forensic Speaker Recognition Evaluation held in 2003 are reported. The goal of this project was to gain an insight in what conditions have to be met for state-of-the-art speaker verification methodologies to be usable in this and similar identification tasks in day to day police work. The speech material used in this evaluation was taken from wire-tapped recordings from real police investigations in the Netherlands. In total six experiments were carried out, one main experiment in Dutch, one experiment in which speech lengths and the number of training sessions war systematically varied, three language dependence experiments, and one experiment evaluating a proposed forensic procedure for providing evidence in court cases. The project found 12 partners who were willing to participate in the evaluation representing the state of the art in current speaker recognition technology. The evaluation was modelled after the yearly evaluations organized by NIST [1], as a one speaker detection task. The lowest Equal Error Rate of all systems was 12.1 % in the condition using 15 seconds test segments and 60 seconds training segments.

Reference:

[1] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. *Speech Communication*, 31:225–254, 2000.

J.S. Bouten is with the Netherlands Forensic Institute (Ministry of Justice), Department of Digital Technology, Rijswijk, The Netherlands.

Paper accepted (yes/no):

Paper number 3

Between-Speaker Variation in Formant Dynamics Associated with Intervocalic /r/

Kirsty McDougall

Formant frequency dynamics are relevant to forensic speaker identification since they are determined by the shape and size of a speaker's vocal tract and the way he or she configures the articulators for speech. Recent studies have noted speaker-specific variation in formant contours, but further work is required to develop techniques utilising this variation for speaker identification. This study examines between-speaker differences in the formant dynamics associated with intervocalic /r/, and investigates ways of using this information to characterise the speakers.

The study focuses on /r/ because realisation of this phoneme has been frequently demonstrated to vary widely between speakers. In particular, speakers can produce the same percept of /r/ via differing articulatory strategies, e.g. bunching versus retroflex approaches combined with varying degrees of lip rounding. If each individual adopts a different articulatory behaviour to achieve the /r/ percept, it is likely that the formant trajectories approaching and leaving the target will yield speaker-specific variation. Intervocalic /r/ was chosen for this investigation so that formant contours would be continuous either side of /r/, and also to enable examination of between-speaker variation in coarticulatory effects on /r/ of adjacent vowels.

Recordings of a group of male native Standard Southern British English speakers reading sentences containing /r/ in a variety of intervocalic contexts are analysed. Centre frequencies of F1, F2 and F3 are measured at time-normalised intervals along the contours of each vowel-/r/-vowel sequence. Attempts to characterise the speakers' formant contours will be made using techniques such as discriminant analysis.Results will be discussed in terms of their applicability to forensic speaker identification. This work also has important implications for speech synthesis, speech recognition systems, and theories of speech production.

Kirsty McDougall is with the Department of Linguistics, University of Cambridge, Cambridge, United Kingdom.

Paper accepted (yes/no):

Paper number 4

Formant Analysis in Forensic Speaker Identification: a Justification and a New Method

Francis Nolan and Catalin Grigoras

Views differ on how useful and how reliable formant measurements are in forensic speaker identification. The majority of cases involve telephone speech, where the restricted bandwidth of the telephone limits the number of formants which can be reliably measured. However it is argued here that formants, whose frequencies and dynamics are the product of the interaction of an individual vocal tract with the idiosyncratic articulatory gestures needed to achieve linguistically agreed targets, are so central to speaker identity that they must play a crucial role in speaker identification. As a practical demonstration a case is described in which F2 analysis of four diphthongs and F1, F2 analysis of a vowel showed a consistent separation between two recordings, thus eliminating a suspect from having made obscene telephone calls. Subsequent additional analysis, based on the long-term frequency distribution of LPC formant estimates, confirms the distinctness of the voice of the suspect and that of the obscene caller.

Francis Nolan is with the Department of Linguistics, University of Cambridge, Cambridge,UK.
Catalin Grigoras is with the National Institute of Forensic Expertise, Bucharest, Romania.

Paper accepted (yes/no):

Paper number 5

Preliminary Observations on Speaker Identification in a Closed Set of Disguised Voices Using LTAS (Long-Time-Average-Spectrum)

Jonas Lindh

Speaker Identification using Graphic Representations of LTAS

Many studies of automatic speaker recognition have investigated which parameters that perform best. This paper presents an experiment where graphic representations of LTAS (Long Time Average Spectrum) were used to identify speakers from a closed set of suspects. Eight different speakers were recorded uttering a fake threat. The utterance durations were approximately 4-5 seconds. The speakers used different disguises such as dialect, accent, whisper, falsetto etc. and the verbatim "threat" in a normal voice. LTAS of the disguised samples were then compared to the "normal" samples to see whether it was possible to identify each disguised voice within a closed set of suspects using only this parameter. No auditory or other approaches were taken into consideration.

The "vocal tract" function in Praat represents LTAS as a graphic line for comparison, not taking the absolute amplitude values into consideration. This is reasonable since the overall absolute amplitude as a parameter has no real value. The important information lies in the relative spectral envelopes represented by a line showing the energy distribution in frequency.

The preliminary results using high quality recordings show a promising performance for the Praat "vocal tract" representation of LTAS. However, more tests on telephone quality recordings and authentic material have to be done to evaluate the use of this method. A comparison between a purely aural approach and the use of this parameter will also be performed.

Jonas Lindh is with the Department of Linguistics, Göteborg University, Sweden.

Paper accepted (yes/no)

Paper number 6

Faking an Accent: The Speaker's Perspective

Niklas Torstensson

There are several possible situations in which a perpetrator would want to disguise his voice in order to avoid identification. Strategies for voice disguise vary greatly, but include imitation or faking a foreign accent. This area has been studied for a number of languages, most often with a focus on the listener and the speaker identification task. This paper presents a study on voice disguise using a foreign accent from the speaker's perspective.

A number of native speakers of Swedish have been recorded, reading a text in three different manners; The first recording where unprepared subjects were asked to read a text in Swedish "as if they were to convince an imagined listener that they were English speakers" followed by a second recording in the subjects' natural voice and reading style. Each subject was then given a recording of a native English speaker reading the same text, and the instruction to listen to it for a week. A third similar recording with accented speech was done after a week.

Both acoustic and auditory comparisons have been made to find intra- and inter-speaker variation over time. The first, unprepared recording reveals intuitive, basic patterns believed to be cognitively prototypical of this certain accent. Inter-speaker comparisons show clear similarities in strategy, but also differences. Intra-speaker variations show some learning effects and changes in production strategy. These variations and changes are discussed and compared with research in the field of voice disguise and voice identification.

These findings contribute to knowledge on speaker strategies in the area of voice disguise, and to cognitive aspects on foreign accented speech in a forensic phonetic context.

Niklas Torstensson is with the Department of Philosophy and Linguistics, Umeå University, Sweden and Department of Humanities and Communication, University of Skövde, Sweden.

Paper accepted (yes/no):

Spectral Change over Time (F1-F3) of Groups of Glide and Vowel in a Polish Dialect (Malopolonian) under Different Communication Conditions – Application in Forensic Cases

Agata Trawińska

The aim of this study was to investigate the spectral change over time (F1–F3) for groups of glide and vowel following the palatalised consonants, especially the bilabial stops, of the Malopolonian Dialect under different communication conditions. It is assumed that the investigated groups of vocoids allow to differentiate speakers clearly and to determine their geographical origins. This research was undertaken, since different degrees of non-synchronically articulated palatality were observed in forensic recordings. The common expression of non-synchronically articulated palatality in Polish is an appearance of glide /j/ (Jassem 1973). Currently, besides the voiced or voiceless inter-words articulation, non-synchronically pronounced palatality are pointed out as the most salient features connected with regional variety in Polish (Urbańczyk 1984).

Method:
Speech material for 12 male Malopolonian Dialect speakers was analysed. From each person, a reading list of words and utterances and repeated utterances including the palatal or palatalised consonants in a phonetically similar context were collected.

	hard consonant	palatalised and palatal counterparts
words and utterances (read vs. repeated)	/a/, /e/, /o/	/ja/, /je/, /jo/

The formants frequency was extracted using the acoustic workstation STx. Groups of vocoids, i.e. glide and vowel such as /a/, /o/ and /e/ was described by tracking of formants (F1–F3) and compared to the structure of formants of the appropriate vowels placed in a phonetically similar context. The way of analysis of the groups of vocoids will be modelled on the description presented in works upon diphthongs in Austrian German (see e.g. Moosmüller 1997).

Reference:

Jassem W., (1973) ‘Podstawy fonetyki akustycznej’ (in Polish).

Moosmüller S., (1997) ‘Phonological variation in speaker identification’, *Forensic Linguistics 4 (1),* pp. 1350-1771.

Urbańczyk S., (1984) ‘Zarys dialektologii polskiej’ (in Polish).

Agata Trawińska is with the Institute of Forensic Research, Cracow, Poland.

Paper accepted (yes/no):

Paper number 8

Fighting the Confirmation Bias: Blind Grouping

Tina Cambier-Langeveld and Erik Jan van der Torre

As a reaction to recent criticism on forensic identification disciplines expressed in for example Saks (1998) and Risinger et al. (2002), a method is being developed at the Netherlands Forensic Institute (NFI) in which information that could bias the findings of the speaker identification expert is banned from the analysis. This method is not designed to replace the standard auditory-phonetic approach used by the NFI, but to supplement it.

The method is applicable primarily in cases where the questioned material comes from wire-tapped telephone conversations. Both the questioned conversations and the reference recordings (preferably telephone conversations as well) are edited in such a way that edit files contain only speech material from either the speaker under investigation or his speaking partner. References to names are taken out. A selection of these edit files are put into one directory and renamed as 1, 2, 3, etc. Another analyst, with no prior knowledge of the case, groups the edit files into groups, without knowing how many speakers may be involved, or how many and which edits contain speech material that has earlier been assigned to the speaker under investigation by the police.

In the presentation, we will discuss the interpretation of the resulting blind grouping and its relation to the results of the non-blind analysis, the reasons for using only speaking partners from the case as 'foils', and some other issues that are brought up by our experience with this method so far.

References:

Risinger, D.M., Saks, M.J., Thompson, W.C. & Rosenthal, R. (2002), 'The *Daubert/Kumho* Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion', *California Law Review* 90(1), p. 1-56.

Saks, M.J. (1998), 'Merlin and Solomon: Lessons from the Law's Formative Encounters with Forensic Identification Science', *Hastings Law Journal* 49(4), 1069-1141.

Tina Cambier-Langeveld is with the Netherlands Forensic Institute, Rijswijk, The Netherlands.
Erik Jan van der Torre is with the Netherlands Forensic Institute, Rijswijk, The Netherlands.

Paper accepted (yes/no):

Paper number 9

Can Spectral Moments Have Perceptual Significance?

Eriksson, E., Cepeda, L., Rodman, R.D., McAllister, D., Bitzer, D., Arroway, P., Sullivan, K.P.H., Sjöström, M., Landgren, T. and Zetterholm, E.

Rodman et al. (2000) proposed a method using spectral moments for forensic speaker identification. This method uses pitch synchronous tracks that are plotted onto a two-dimensional plane. The plane is comprised of two moments plotted against each other. The moments are the mean and variance about the mean taken from the probability density function from a spectrum. The spectrum in turn is taken from similar sounding segments of a speech signal. The two moments together forms a track. The algorithm is presented in Rodman et al. (2000). The positions of the tracks on the two-dimensional plane have been shown to be speaker dependent (Rodman et al 2000, Eriksson et al, in press).

A voice line-up experiment with imitated speech of a famous Swedish politician showed that listeners frequently choose another voice than either the imitators. natural voice or the real voice of the politician (Zetterholm et al. 2003, Eriksson et al. 2003). Non-native speakers have showed no such confusion (Sullivan et al. 2002, Zetterholm et al. 2002). Several factors have been investigated without leading to a satisfying explanation. By utilizing the spectral moments and visually displaying the tracks on the aforementioned two-dimensional plane, it can be seen how close the tracks are to each other. If the tracks are close to each other and similar in structure, it can be said that the method captures aspects of the cognitive representation of human voice perception. This being the case, the distance between tracks in the two-dimensional plane ought to correlate with the confusions made by the listeners.

We begin the paper by presenting an overview of the spectral moments approach, and then presenting the results of the perception studies. Thereafter the results of analyzing the same speech using the spectral moments approach is presented and its success rate is considered from both a cognitive and a mathematical perspective.

References:

Eriksson, E., Kügler, F., Sullivan, K.P.H., van Doorn, J. and Zetterholm, E. (2003). Why foil 4? A first look. Phonum 9: 161-164. Proceedings from Fonetik 2003, Umeå, June 2 – 4 2003.

Eriksson, E., Cepeda, L., Rodman, R.D., McAllister, D., Bitzer, D. and Arroway, P. (in press). Cross-language speaker identification using spectral moments. Proceedings from Fonetik 2004, Stockholm, May 26 – 28 2004.

Rodman, R.D., McAllister, D., Bitzer, D., Cepeda, L. and Abbitt, P. (2002). Forensic Speaker Identification based on Specral Moments. Forensic Linguistics 9(1): 22 – 43.

Sullivan, K.P.H., Zetterholm, E., van Doorn, J., Green, J., Kügler, F. and Eriksson, E. (2002). The effect of removing semantic information upon the impact of voice imitation. Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, December 2 – 5 2002.

Zetterholm, E., Sullivan, K.P.H. and van Doorn, J. (2002) The impact of semantic expectation on the acceptance of a voice imitation. Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, December 2 – 5 2002.

Zetterholm, E., Sullivan, K.P.H., Green, J., Eriksson, E., van Doorn, J. and Czigler, P.E. (2003). Who knows Carl Bildt? – And what if you don't? Eurospeech 2003 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, September 1 – 4, 2003.

Paper accepted (yes/no):

Inter- and Intra-Speaker Variation in F0 Parameters – Twin Speech

Leena Keinänen and Tuija Niemi-Laitinen

The purpose of this study is to find out the differences and similarities between the speech of monozygotic and dizygotic twin pairs. Especial interest is put on prosodic features in speech. These features include speaking fundamental frequency, intonation, duration of speech segments, stress, rhythm as well as speech and articulation rate. In this study we concentrate on speaking fundamental frequency and its different parameters. The speakers are 18 female twin pairs (9 MZ, 9 DZ), ages from 22-28 years. The speech data includes both reading a story and telling spontaneously from a picture. The speech data has been recorded via microphone and saved to a database. Praat program has been used for the measurements. A special script for measuring the F0 parameters has been developed for Praat. This script analyses minimum, maximum, mean, median and standard deviation values of F0.

A common belief has been that the speech of monozygotic twins is very similar and the differences inside the twin pair is smaller compared to the differences inside a dizygotic twin pair.

The results of this study show that there is a lot of variation between the monozygotic twin pairs' speech. This variation is sometimes greater than with the dizygotic twin pairs.

The details of the results will be discussed in the final paper.

Leena Keinänen is with the Department of Phonetics, University of Helsinki, Finland.
Tuija Niemi-Laitinen is with the Crime Laboratory, National Bureau of Investigation, Finland.

Paper accepted (yes/no):

Paper number 11

Acoustical Differences between Mobile Phones

Tuija Niemi-Laitinen and Kirsi Harinen

GSM speech material is very common at the Speech and Audio Analysis Section in the Crime Laboratory. For quality reasons it is necessary to know what kind of differences can occur between different GSM phone types.

The aim of this study is to find out acoustical differences in some speech parameters that certain GSM phones produce. The speech material used in this study consists of read speech. Two speakers, one male and one female, read a 3 minute long passage twice. The recordings were made at the Dept. of Phonetics, University of Helsinki with a DAT recorder and a good quality microphone in a quiet room. The speech material was stored in a computer in 24 bit resolution at sample rate 44,1 kHz. These recordings were played via loudspeakers to different kinds of GSM phones. One phone, Nokia 3330 was used twice in order to find out whether there is any effect of random channel connection. The other phone types were Nokia 6100, Nokia 5110, Siemens A50, Bosch DUAL – COM738 EFR, Benefon, and one land-line phone (Siemens Euroset 815S). These phones were hold in a distance of 5-7 cm apart from the loudspeaker during the sessions. The receiving apparatus was situated in the Crime Laboratory where the speech was stored into the computer's memory at 44 kHz, 16 -bit wav files.

The recordings of the same speech material were used with every phone type. This was made to avoid any intra-individual differences that can occur between different reading sessions.

The parameters that were used in this study were LTAS and speaking fundamental frequency (F0 mean, median, minimum, maximum and standard deviation). The results of this study are used in method development of the Crime Laboratory and the estimation of the reliability of the statements.

Tuija Niemi-Laitinen is with the Crime Laboratory, National Bureau of Investigation, Finland.
Kirsi Harinen is with the Department of Phonetics, University of Helsinki, Finland.

Paper accepted (yes/no):

Paper number 12

The Effect of Compression on Speech Intelligibility

Päivikki Eskelinen-Rönkä and Tuija Niemi-Laitinen

Speech compression is a fairly new and difficult area in forensic studies, and it has naturally consequences in reliability of both quantitative and qualitative analysis. The question of intelligibility is specifically vital in those cases where transcription analysis is necessary.

The focus of this study is to find out the possible effects of three different compression algorithms on speech intelligibility. The compressions under investigation are options in Nice Log –archiving system that is widely used in Finland, for example in emergency centres and in air traffic control units.

Speech material used in listening tests consisted of authentic aviation communication recordings that were compressed by PCM 64, ADPCM 16 and ACA 5.6 algorithms. Listeners were divided in two groups according their experience in aviation radiotelephony communication: the professionals and the naives. The first listener group was called "professionals" and it consisted of pilots and air traffic controllers. The second group of listeners was called "naives" and these listeners did not have any previous experience of aviation communication.

The purpose of the listening test for the professionals was to find out whether the compression algorithms had a tendency to produce predictable errors in this kind of speech material. It was also important to find out how compression affects intelligibility of speech in those cases where the listener is naive and not familiar with the speech content that has to be transcribed. Another question was how different compressions affect intelligibility of speech when the language in question is different from the native language of the listener.

Päivikki Eskelinen-Rönkä is with the Department of Phonetics, University of Helsinki, Finland.
Tuija Niemi-Laitinen is with the Crime Laboratory, National Bureau of Investigation, Finland.

Paper accepted (yes/no):

Paper number 13

Automatic Speaker Identification in Forensic Analysis; How Useful Is It?

Mira Wedemeyer

A problem often encountered in connection with forensic investigations is the identification of a criminal or witness by an excerpt of his voice. Naturally, it would be useful to have an automatic speaker identification via computer. During the last years, more and more research has been done on this topic, as there are many useful applications for speaker identification aside from forensic analysis, for example the recognition of a news anchorman for an automatic metadata extraction. Speaker identification is split into two main challenges. The first challenge is the extraction of features that may represent the speaker. Commonly used features are the Mel Frequency Cepstral Coefficents (MFCC's), and the Audio Spectrum Envelope (ASE), which is defined in the MPEG-7 standards. Both extraction methods work in the frequency domain, and are based on the acoustical properties of the human ear. Apart from these methods, delta coefficients are often used to characterize a speaker. These coefficients represent the temporal changes of the above-mentioned features. The second challenge in speaker identification is assigning the features to a speaker. This challenge is encountered by methods such as Vector Quantisation (VQ), or Gaussian Mixture Model classification (GMM). In this contribution the aforementioned methods for speaker identification are introduced. Following this, the benefits of these methods for forensic analysis are investigated. In forensic analysis, speech signals are often highly distorted so that automatic speaker identification may fail. The causes of these failures are shown and a final evaluation of the benefits of speaker identification in forensic analysis is given.

Mira Wedemeyer is with the Houpert Digital Audio, Bremen, Germany.

Paper accepted (yes/no):

Paper number 14

The Forensic Analysis of a Possible Misperception

Allen Hirson

A case of alleged racial abuse rested on a claim by a Police Officer that racially abusive words had been shouted at him. The incident took place in the evening on a busy suburban roadside, and the accused (M) had been drinking. Both parties were besides the police vehicle, which had its flashing lights on. M's responses to most questions at interview the same evening were transcribed by the police as 'inaudible'. M subsequently claimed that he had made no derogatory comments, but had shouted 'Turn the fucking lights off!' as he was having difficulty putting his 2 year-old grand-daughter to sleep within his house. Mr M was described by his lawyer as deaf and somewhat unintelligible.

Data was collected regarding noise and lighting at the site of the incident, M's audiological status and his speech intelligibility. Hearing was assessed by pure tone audiometry, sound level measurements were made of the traffic noise at the site, and informal assessements were made of roadside lighting and of M's overall communicative competence. Speech intelligibility was measured using the QuickSIN: Speech in noise Test, Etymotic Research, 2001, which is based on listener transcriptions of test words embedded in carrier sentences. These data were recorded to DAT and MiniDV and transcribed by normally hearing listeners (N=8, 4 males; 4 females). M's data and matched material from a control (normally hearing) speaker were also evaluated by two audiologists with experience of deaf speech using a standard rating scale (Cox & McDaniel, 1989).

Unexpected intelligibility scores from the video recordings are discussed in relation to M's effortful speech and also to the reliability of lip-reading evidence (Campbell-Tiech, 2002). Possible effects of alcohol (and shouting) on the speech intelligibility are also discussed. Conclusions regarding the /reliability/ of the claim of racial abuse led the Crown to drop the case.

References:

Campbell-Tiech A. (2002) Lip reading as expert evidence. /Archbold News/, 2002, 5, 5-6.

Cox R. M. & McDaniel D. M. (1989) Development of the Speech Intelligibility Rating (SIR) test for hearing aid comparisons/. Journal of Speech and Hearing Research/, *32*, pp. 347-352.

Lester L. & Skousen R. (1974) The Phonology of Drunkenness. In /Papers from the Parasession on Natural Phonology/, Chicago Linguistic Society, April 1974, pp. 233-239 (eds. Bruck, A., Fox R., La Galy, M. W.).

Allen Hirson is with the City University, London, England, UK.

Paper accepted (yes/no):

Paper number 15

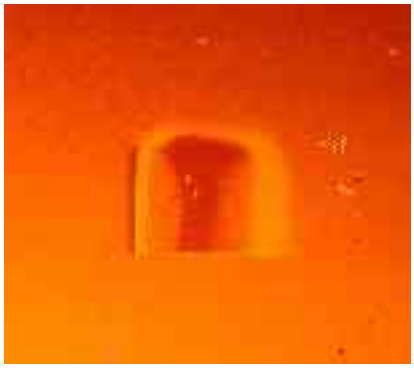
Authentication of an Evidence Tape by Visualization of Magnetic Features

Dagmar Boss

Within the field of tape authentication the analysis of start and stop events usually plays an important role. When the “record” function of an analogue recording device is turned on or off, both record and erase heads (or a combined head) produce a mark on the tape, the erase head mark being the more interesting one of them. These marks can be investigated best by using not only waveform and spectrographic analysis but also the visualization of magnetic features, which can be carried out easily today by using special, artificially produced crystals.

In the criminal case to be presented some important questions concerning an evidence tape could be solved by means of visualization. The main question concerning the tape was here whether it had been stopped one or more times during the recording or not (if the tape had not been stopped, its content clearly would help to discharge the defendant). This question had arisen mainly because of the fact that there were 60 signal interruptions within the three minutes of conversation. Most of them were only about 100 ms long. Their origin was completely unclear.

All the interruptions were investigated thoroughly, first by waveform and spectrographic analysis, and then by visualization. The visualization could show that the marks related to the interruptions had been produced by two different devices: the very short ones (57 out of 60) by the recorder on which the tape was listened to by the person who had to write down the text. The record-enable tabs had not been removed before, so that a simple handling mistake (pushing the “record” instead of the “pause” button) left its marks on the tape. This relationship between the marks on the evidence tape and the device used for listening could be established only by visualization: the marks at almost all the interruptions showed the same pecularities as the marks which were produced on the same device for comparison reasons.



One of the marks

These pecularities could not be shown by waveform or spectrographic analysis.

The mishandling of the recorder explained only 57 out of 60 interruptions – the other three ones looked very much as if the recording really had been stopped and re-started there. In addition to that, the magnetic features clearly showed that these three stops were produced on another device than the interruptions that could be related to the mishandling. They looked very similar to marks within other recordings which were made on the same device as the incriminated conversation. Thus, it could be considered as probable that during the recording of the conversation, the tape had been stopped and re-started three times.

Dagmar Boss is with the Bayerisches Landeskriminalamt, München, Germany.

Paper accepted (yes/no):

Paper number 16

Forensic Formant Frequency Measurements: Issues and Instrumentation

Michael Jessen

According to the Acoustic Theory of Speech Production (Fant, Stevens etc.) speakers with a short vocal tract show a shift of the vowel space towards higher frequencies relative to speakers with a longer vocal tract. Since different individuals can differ in vocal tract length, the speaker-specific value of formant frequencies is predicted. This prediction is borne out by a large body of literature. In this contribution theoretical and practical aspects of the measurement and interpretation of formant frequencies in forensic speaker identification will be discussed. This study is based on a survey of the literature and on casework data. The presentation includes the following topics:

* Introduction: why measure formant frequencies? (high speaker specificity; independence and complementary advantages of formant and f0 measurements; small telephone channel influence etc.)

* Formant frequency measurement methodology: experience with S_TOOLS-Stx (LPC formant tracking and correction, labeling and segment administration, data export etc.)

* The different functions of formant frequencies (information on / influence by anatomy, tempo / articulatory precision, dialect, setting, coarticulation etc.)

* Manipulating formants with commercial software

* Case studies

Michael Jessen is with the Department of Speaker Identification and Tape Analysis, Bundeskriminalamt, Germany.

Paper accepted (yes/no):

Paper number 17

On the Use of Auditory and Automatic Systems to Handle Mismatched Conditions in Forensic Speaker Recognition

Anil Alexander, Damien Dessimoz, Filippo Botti and Andrzej Drygajlo

In this paper we analyse mismatched technical conditions in training and testing phases and their effect on human and automatic forensic speaker recognition. Automatic speaker recognition has shown high performance under controlled conditions. The recording conditions in which recordings are made by the police in forensic cases (anonymous calls and wiretapping) cannot be completely controlled. Differences in the phone handset, in the transmission channel and in the recording tools introduce a variability, over and above the variability of human speech. We use perceptual tests performed by non experts and compare their performance with that of an automatic speaker recognition system. These experiments are performed on 60 phonetically untrained subjects. Several forensic cases were simulated, using the IPSC02 Polyphone database, varying in linguistic content and technical conditions of recording. We estimate the strength of evidence for both humans and the baseline automatic system, calculating likelihood ratios using the perceptual scores for humans and the log-likelihood scores for the automatic systems. An analogous methodology to the Bayesian interpretation in forensic automatic speaker recognition is applied to the perceptive scores given by humans in order to estimate the strength of evidence. The degradation of the accuracy of human recognition in mismatched recording conditions is contrasted with that of the automatic system under similar recording conditions. The conditions considered are fixed telephone, cellular telephone and noisy speech. The perceptual cues that the human subjects use to perceive differences in voices are studied along with their importance in different conditions. We discuss the possibility of increasing the accuracy of automatic systems using the perceptual cues that remain robust to mismatched conditions.

Anil Alexander is with the Signal Processing Institute, Swiss Federal Institute of Technology (EPFL).
Damien Dessimoz is with the Institut de Police Scientifique, University of Lausanne (UNIL), Lausanne, Switzerland.
Filippo Botti is with the Institut de Police Scientifique, University of Lausanne (UNIL), Lausanne, Switzerland.
Andrzej Drygajlo is with the Signal Processing Institute, Swiss Federal Institute of Technology (EPFL).

Paper accepted (yes/no):

Paper number 18

The Speaker Discriminating Power of the Final Fall

Hazel Walters, Gea de Jong and Jill House

Several studies have found that, in different languages, the 'final' F0 value of utterance-final simple falls is a relatively invariant characteristic of a speaker's voice (British English: Nolan 2002, American English: Liberman and Pierrehumbert 1984, German: Oppenrieder 1988, Dutch: Kraayeveld 1997, Mexican Spanish: Prieto et al. 1996). The measurements reported in a majority of these studies were based on utterances of similar linguistic content where the final syllable ended in a vowel or all sonorant codas. Some studies focused on the absolute terminal F0 value, while others measured the lowest F0 within the falling pattern, or the contour 'elbow', preceding the final F0. The aim of this study was to extend the above findings by 1) investigating whether final F0 values belonging to utterance-final falls continue to be a relatively stable feature when applied to syllable codas of varying phonetic content, 2) comparing the final fall means for a series of different sentences, 3) comparing the discriminative power of the F0Final with the power of the F0Lowest parameter, and 4) investigating whether the small variations in the physical shape of the F0 contour-Fall type were related to linguistic content or individual preference. Recordings of 10 female and 10 male speakers (Markham&Hazan, UCL-database, 2002) were studied saying the same 20 semantically unpredictable sentences (Benoit, Grice&Hazan, 1996).

Results showed that:

1) For both F0Lowest and F0Final, the within-speaker variability is significantly smaller than the between-speaker variability,

2) F0Lowest had a greater discriminatory power than F0Final, with F0Lowest significantly discriminating 11 subsets out of the entire set of 20 speakers at $p<0.05$ and F0Final discriminating 8 subsets, and

3) 'Fall-type' was more dependent on individual preference than coda structure.

Hazel Walters is with the Phonetics and Linguistics Department of University College London, UK.
Jill House is with the Phonetics and Linguistics Department of University College London, UK.
Gea de Jong is with the Language and Communication Science Department of City University, London, UK.

Paper accepted (yes/no):

Paper number 19

One Speaker: Two Voices — One Imitator: Two Voices

Elisabeth Zetterholm and Kirk P.H. Sullivan

A listener's cognitive capacity to remember voices and identify them is dependant upon many factors. The voice selected in a voice line-up has been shown to be affected by expectation. Expectation can include familiarity with the target voice and the topic of the passage presented in the voice line-up. The success of voice imitation as a factor in voice line-up settings is dependant upon such factors. This paper investigates one male Swedish speaker who is bi-dialectal and voice imitations of his 'two voices' by one professional impersonator. The impersonator was not informed that his task was to imitate passages spoken by a single speaker, but rather a set of passages spoken by two different speakers. The accents spoken by the speaker were the accent of Stockholm and that of Skania (Southern Sweden). These accents differ markedly both prosodically and segmentally and the expectation that they are spoken by one person (and accepted as a native speaker of both dialects) is minimal. The two voices of a bi-dialectal person, thus, afford the opportunity to examine natural within speaker variation, the features that are the same for both voices and those that are different in the imitations, and listeners' ability to recognize cross-dialectal voices in line-up contexts.

The data presented represents the initial stage of the investigation. They include an auditory analysis of the voices (that is the 'four voices'), a small set of general acoustic measures and the results of a sequence of voice line-ups experiments. These voice line-ups do not include the imitations but vary the bi-dialectal voices included. The experiments show that the speaker is not recognised when speaking the accent that is not expected by the listener.

Elisabeth Zetterholm is with the Umeå University, Sweden.
Kirk P.H. Sullivan is with the Umeå University, Sweden.

Paper accepted (yes/no):

Paper number 20

Sociolinguistic and Acoustic Variability in Filled Pauses

Paul Foulkes, Gareth Carrol and Samantha Hughes

At the 2003 IAFPA conference in Vienna we presented preliminary results from an acoustic study of filled pauses (*uh* and *um*) in English. Thanks to a research grant from the IAFPA we have now completed this study and will present definitive results at the Helsinki conference.

It has been suggested that filled pauses may be good diagnostics in forensic speaker identification. It has been claimed, for example, that individual speakers remain relatively consistent in the acoustic qualities they use for the vocalic portions of *uh* and *um* (e.g. Künzel 1987, Pätzold & Simpson 1995). Vocalic sounds used in linguistic utterances, by contrast, are subject to a wide range of factors which can induce variability, including phonologically-conditioned allophony, sociolinguistic and stylistic factors, coarticulatory effects of phonological context, and speech rate effects. The more consistent patterns found for the vocalic portions of filled pauses therefore offer a potentially good cue to speaker identity.

Our study tests such claims using a large sociolinguistically-balanced sample of speakers. The data were extracted from the Newcastle recordings of the *Phonological Variation and Change* project (Milroy et al 1997). Over 1,200 tokens of *uh* and *um* were analysed from 32 speakers, sampled along age, class and sex dimensions. Using the Praat program values of the first three formants were taken, and compared with reference vowels from the same speakers. Our analyses suggest that filled pauses are marginally less variable than lexical vowels. F1 and F2 measurements of filled pauses were for most speakers no less variable than lexical vowels. F3 measures, however, were less variable in 40% of comparisons. The formant data were also used in discrimination tests. It was found that the filled pause data performed as well as, or marginally better than, data drawn from lexical vowels.

References:

Künzel, H. (1987) *Sprechererkennung: Grundzüge Forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag.

Milroy, L., Milroy, J. & Docherty, G.J. (1997) *Phonological Variation and Change in Contemporary Spoken British English*. Final Report to the UK Economic and Social Research Council, grant no. R000234892.

Pätzold, M. & Simpson, A. (1995) An acoustic analysis of hesitation particles in German. *Proceedings of the 13th International Congress of Phonetic Sciences*. Stockholm. Vol. 3. pp. 512-515.

Paul Foulkes is with the University of York, UK.
Gareth Carrol is with the University of York, UK.
Samantha Hughes is with the University of York, UK.

Paper accepted (yes/no):

Paper number 21

On Likelihoods, Probabilities and Uncertainty in Forensic Automatic Speaker Recognition

Andrzej Drygajlo

The interpretation of recorded speech as evidence in the forensic context presents particular challenges. The means proposed for dealing with this is through Bayesian inference and corpus based methodology. Likelihood-based and probability-based models can be used for assisting forensic experts in the speaker recognition domain to interpret this evidence. This leads to to the formulation of a likelihood ratio and an error ratio measure of evidence, which weighs the evidence in favor of two competing hypotheses: 1) suspected speaker is the source of the questioned recording (trace), or 2) he is not. In forensic speaker recognition, statistical modelling techniques are based on the distribution of various features pertaining to the suspect's speech (within-source variability) and its comparison to the distribution of the same features in a reference population (between-sources variability) with respect to the questioned recording.

In order to model multivariate data arising from speech signal we use two principal methods: 1) direct method, which directly uses the likelihoods returned by the Gaussian Mixture Models (GMMs), and 2) scoring method, which models the distributions of these likelihood scores and then derives the likelihood ratio on the basis of these score distributions. In this paper, the automatic, text-independent GMM-based speaker recognition system is adapted to the Bayesian interpretation (BI) framework to estimate the within-source variability of the suspected speaker, the between-sources variability of the potential population, given the questioned recording, and the strength of evidence. Two complementary measures of interpreting the evidence within the Bayesian framework are compared: likelihood ratio and error ratio. The direct and scoring methods can be used when the statistical model of the within-source variability of the suspected speaker's speech is available. If not, a third indirect method has to be introduced, which uses estimated within-source variability of speakers different from the suspected speaker but recorded in similar conditions.

Andrzej Drygajlo is with the Speech Processing and Biometrics Group, Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland.

Paper accepted (yes/no):

Paper number 22

The Effect of Standard Audio Compression Algorithms on Forensically Relevant Measurements

Angelika Braun and Carsten Grasmück

There are various reasons why forensic phonetic experts may be confronted with compressed data. First, telephone tap data are nowadays almost always collected in a compressed format; the same is true for other sorts of clandestine recordings. Also, the quick transmission of forensic data may require audio compression.

While these compression algorithms were expressly developed with the aim of leaving the auditory impression unchanged, they may still affect certain acoustic measurements. Specifically, spectral parameters like formants and spectrum-based F0-measurements are potentially affected. That is why the "net" effect of these algorithms on an "ideal" signal was studied.

A total of 30 speakers (15 male and 15 female) participated in the present study. They were asked to read a standardized text (The North Wind and the Sun), pronounce a list of words containing different vowels of German and describe a cartoon from Max and Moritz by Wilhelm Busch.

The recordings were made using professional equipment. There were then digitally compressed using two different coding algorithms (ATRAC and MP3).

The measurements which were carried out cover typical procedures in forensic phonetics: They included F0 and F0 standard deviation using different algorithms, formant frequencies and formant bandwidths.

Results show some of the measurements to be statistically significant (Wilcoxon test for paired samples), the MP3 coding having a stronger effect than ATRAC. Also, effects were stronger on female voices than on males. However, the magnitude of the differences in absolute terms was always so small that – based on these data – the "net" effect of those two compression algorithms will not have a detrimental effect on forensically relevant measurements. It would be interesting, though, to study the effect of coding on the new generation of semi-automatic SR-systems.

Angelika Braun is with the Universities of Marburg and Trier, Germany.
Carsten Grasmück is with the University of Trier, Germany.

Paper accepted (yes/no):

Paper number 23

Adapting the Praat Speech Analysis Programme to the Purposes of Forensic Phonetic Casework and Research

Peter French and Philip Harrison

This is a practical demonstration of script developments and modifications to the Praat speech analysis programme. The initial motivation for the work was to facilitate casework and research in forensic phonetics. However, the developments have much wider applications within phonetics teaching and research.

The modified version of the programme allows one to track and log the centre frequency values of vowel formants. When the tracking and logging procedure is complete, the programme automatically calculates the following in respect of f1, 2, 3 and 4 for the tokens of each vowel phoneme analysed:

Average frequency
Maximum and minimum values
Range
Standard deviation from the mean

The same information is provided for the delta scores: f2 - f1, f3 - f2 and f3 - f1.

On the basis of their f1 and f2 – f1 delta values, the vowel tokens analysed are also mapped into vowel space on an automatically-generated vowel quadrilateral.

The modifications allow one to generate large amounts of elegantly presented useable data with a small number of keystrokes.

(A brief demonstration of further, non-formant related, adaptations is also included.)

Peter French is with the JP French Associates, York, UK.
Philip Harrison is with the JP French Associates, York, UK.

Paper accepted (yes/no):

Paper number 24

Automatic Language Recognition with Spectral and Prosodic Features

Seppänen T., Väyrynen E., Toivanen J. and Iivonen A.

This paper investigates the usefulness of spectral and prosodic features for statistical classification of speech samples into predetermined languages. Several versions of the recognizer are implemented to test how much added value prosodic features offer to supplement basic cepstral information. Baseline version of recognizer utilizes language-specific codebooks of normalized prototypical short-time cepstra that are formed by a learning vector quantizer algorithm. In short-time cepstra, the window length is chosen to approximate an average phone length, in order to produce approximate spectral models for most frequent phones in the data. When recognizing an unknown sample, the sequence of short-time cepstra collected from an unknown speech sample is matched to the codebooks on a vector-by-vector basis. The vector differences are then combined to calculate the overall distance of the sample from each language model for nearest neighbor classification. The second recognizer is compeletely based on prosodic features. More than 50 features are calculated from each speech sample used for training the recognizer. Normalized feature vectors are collected for each language data. A stepwise forward-backward feature selection algorithm with leave-one-out testing approach and kNN classifier is run over the whole dataset in order to select the most capable features for discriminating the languages. A language model is formed that uses all the samples as represented with these best features. A similar feature vector is then calculated from an unknown speech sample, and a kNN classifier is used for recognizing the language. The third recognizer is a hybrid of these two. In the first stage, a baseline recognizer is applied. If a tie occurs, the prosodically-based recognizer resolves the situation. In our experiments, microphone-level Finnish, Estonian, Russian, French, English, Turkish, and Spanish are used as speech data.

Tapio Seppänen is with the Department of Electrical and Information Engineering, University of Oulu, Finland.
Eero Väyrynen is with the Department of Electrical and Information Engineering, University of Oulu, Finland.
Juhani Toivanen is with the Department of Electrical and Information Engineering, University of Oulu, Finland.
Antti Iivonen is with the Department of Phonetics, University of Helsinki, Finland.

Paper accepted (yes/no):

Paper number 25

Automatic Segmentation and Extraction of Speaker Specific Segmental Cues

Antti Iivonen, Kirsi Harinen and Jussi Kirjavainen

Our automatic SpeakerProfiler based on Praat scripting comprises already robust acoustic features which have been shown to be very effective in speaker identification, if the number of speaker is about 100. We expect that with much greater speaker amount the number of correct identifications might not be as successful. Additional rubust cues are therefore necessary, such ones which could increase the discriminatory power of the profiler. Our next step is based on (1) automatic segmentation of speech , (2) on the fact that the verbal contents of the speech signal is known and (3) that the acoustic signal can be forced to segmentation. A Praat script has been created which uses the automatic segmentation. The feature extraction is concentrated on specific selected segment addresses. So far, we have used [r]-segments of Finnish. The dominant Finnish [r] is a trill which can vary in strongness and in number of vibrations. Further types of speaker specific variation occur: tap, fricative alveolar sound, trilled uvular sound. Futhermore, according to quantity a phonologically short and long categories exist. Smoothed averaged FFT-spectra, the frequency band 0–3400 Hz band (devided in 200 subbands) and a 20 ms long time window have been used. Special attention has been paid on the coarticulatory aspect of [r]. Euclidian distances have been used in order to estimate the intra- and interspeaker variablity and the usability in speaker recognition.

Antti Iivonen is with the Department of Phonetics, University of Helsinki, Finland.
Kirsi Harinen is with the Department of Phonetics, University of Helsinki, Finland.
Jussi Kirjavainen is with the Department of Phonetics, University of Helsinki, Finland.

Paper accepted (yes/no):

Paper number 26

Speaker Identification: A Linguistic Approach

Gea de Jong and Peter Smith

In (Smith and DeJong 2003) we suggested that speaker identification may be improved by a joint phonetic analysis and stylometric analysis of the transcribed text. Customary stylometric techniques are more geared towards higher volumes of text than are usually available in forensic cases. One stylometric analysis technique that can in principle be applied to lower volume text and that has been used in a forensic setting is the cusum technique (Morton and Michaelson 1990). Serious doubts have already been raised about this technique (Robertson et al. 1996), (Hardcastle 1997), but we decided to re-examine it on the problem of speaker identification rather than dismissing it completely. We attempted to follow the principles specified in (Farringdon 1996), but immiediately found serious problems of methodology and were completely unable to establish any form of consistency for this technique even over one reference text in which speaker identification is beyond question. In (Smith and DeJong to be published) we describe our experiment with the Cusum method and explain why the "habits" used for speaker identification with this method are inadequate.

An alternative technique that apears promising is the use of Markov models in authorship attribution. (Khmelev and Tweedie 2001) describe successful attribution over a wide range of texts. However, in a forensic application it is important that the method can be articulated clearly. The method described in (Khmelev and Tweedie 2001) is based on letter collocations and as they freely admit, it is not clear how to condense the Markov tables into a coherent explanation of style - something that would be essential in a legal setting. We believe that this method can be extended to a Markov model analysis of words and even grammatical constructs and are currently working with Part-of-Speech taggers which should enable us to provide an explanation-based analysis. By using grammatical structures and function words, we may be able to produce such an analysis.

Gea de Jong is with the Department of Language and Communication Science, The City University, London, UK.
Peter Smith is with the Department of Computing, The City University, London, UK.

Paper accepted (yes/no):

Paper number 27

Spectral Fusion System for Speaker Recognition

Tomi Kinnunen, Ville Hautamäki and Pasi Fränti

Several features have been proposed for automatic speaker recognition. Despite their noise sensitivity, low-level spectral features are the most popular ones because of their easy computation. Although in principle different spectral representations carry similar information (spectral shape), in practise the different features differ in their performances. For instance, LPC-cepstrum picks more “details” of the short-term spectrum than the FFT-ceptrum with the same number of coefficients. In this work, we consider using multiple spectral presentations simultaneously for improving the accuracy of speaker recognition. We have selected to use the following feature sets: mel-frequency warped cepstral coefficients (MFCC), LPC-cepstrum (LPCC), arcus sine reflection coefficients, formant frequencies, and the corresponding deltaparameters of all feature sets. We study the two ways of combining the feature sets: input-level fusion (feature vector concatenation) and output-level fusion (soft combination of classifier outputs).

Tomi Kinnunen is with the Department of Computer Science, University of Joensuu, Finland.
Ville Hautamäki is with the Department of Computer Science, University of Joensuu, Finland.
Pasi Fränti is with the Department of Computer Science, University of Joensuu, Finland.

Paper accepted (yes/no):

Paper number 28

The Forensic Speaker Recognition: a Methodological Comparison

Laura Mori and Andrea Paoloni

The present study proposes a comparison between two speaker recognition systems: the semi-authomatic system IDEM in use at Fondazione Ugo Bordoni for forensic speaker recognition and an authomatic speaker recognition system based on Mel-cepstrum parameters.

As far as IDEM is concerned, this system is based on phonetic measures (formantic values and AR values) which are statistically treated in order to evaluate the identification rate of different voices.

On the other hand the ASR system calculates the distance distribution between couples of samples by different speakers allowing to have a comparison rate among different voices.

Our objective is to compare the identification results obtained by using either an authomatic or a semi-authomatic speaker recognition system. For this purpose we led an experiment using data extracted by a forensic corpus of phonic expertises where the recognition task was to recognize among speech samples from anonymous recordings and several speech samples produced by under suspicion' persons people (repetition of selected utterances).

The result obtained with this methodological comparison allowed us to underline that the identification task led by using a semi-authomatic system such as IDEM gives a greater separation between different speakers.

Laura Mori is with the Università degli Studi di Roma 'La Sapienza', Viterbo, Italy.
Andrea Paoloni is with the Fondazione Ugo Bordoni, Roma, Italy.

Paper accepted (yes/no):

Comparison of Speech Characteristics of Male Voice Distorted by Technical Devices with Its Standard Characteristics (on the Example of "The Voice Changer-2")

Rodmonga K. Potapova and Vsevolod V. Potapov

It is possible nowadays to change the voice of a participant of a conversation with the use of the latest technologies. Many modern programs of analysis and conversion of audio signals include pitch shifting and voice changing without the change of the speech tempo. About ten firms in the USA produce special devices that can be connected to the telephone and contain a microchip that is made to change the voice of a speaker. Families who have small children or elderly or ill people who often have to stay alone use them. But the thing is that the voice changing systems are being widely used by criminals and it is very difficult to discover this fact. That's why the subject of this research work is justified: one should examine the features of the devices of this type to state the fact of the changing and to try to identify the changed voice with the original.

We touch upon the production of special devices for voice changing and the ability to identify a person in the cases when the voice was changed. An experiment to analyze the effects of the use of voice changing or pitch-shifting algorithm was carried out on the example of "The Voice Changer-2". It electronically re-structures human voices to have real-time telephone conversations in a disguised voice. The device is made to offer 8 different voice settings, ranging from the so-called child's voice to a very low adult's voice with natural sounding speech.

Material under test: the voice of a male speaker (original and changed). Procedures: the analysis of the changes between the standard acoustic signal of male voice and the one that was distorted by "The Voice Changer-2": Pitch analysis, length and formant analysis.

Devices/programs used: a Panasonic KXFX 130 BX telephone with a built-in tape recorder; CSL 4300 a digital converter specially made for personal computers; The Voice Canger-2/Multi-Speech Analysis Workstation, model 3700-16-bit, version 1.20; Sound Forge Cool Edit

Rodmonga K. Potapova is with the Moscow State Linguistic University, Moscow, Russia.
Vsevolod V. Potapov is with the Moscow State Lomonosov University, Moscow, Russia.

Paper accepted (yes/no):

Paper number 30

Quantitative Analysis in Speaker Identification

Kangsheng Li and Zhiqiang Xiong

Vowel is nuclear of syllable in Chinese dialects. In this study, we use a new method of vowel (vowel nuclear) identification by adopting multivariate statistical analysis. By analysis on coefficient of variation, coefficient of correlation, test for normality, and analysis of variances (ANOV), we construct a identification model of Mahalanobis distance method for male speakers with the correct rate of 98.3%, and a model of Fisher's discriminant method for female with the correct rate of 93.5%.

Keywords: Parameters of acoustic cues, Multivariate statistical analysis, Speaker identification, Mahalanobis distance, Fisher's discriminant

Kangsheng Li is with the Shenzhen Prosecution service, China.
Zhiqiang Xiong is with the Shenzhen Prosecution service, China.

Paper accepted (yes/no):

Paper number 31

Telephone Speaker Identification within a Family Group

Elizabeth McClelland

This study sets out to test the validity of the common assumption, shared by police officers and barristers, that people can accurately identify over the telephone the voices of individuals that they know. The subjects were selected on the basis that they had frequent face-to-face and telephone contact over a long period and would be expected, if this assumption is correct, to achieve a high level of accuracy in identification.

A group of fourteen close family members, aged 17 – 85, consisting of seven males (including one set of twins) and seven females, were asked to make two telephone calls, one on a landline, the other on a mobile telephone. They were asked to leave one message in which they read from a phonetically balanced text, the other was a loosely scripted message of a domestic nature. Extracts of the messages were played to the subjects who had made the calls and they were asked to identify the voices. The results were assessed in terms of the transmission medium and linguistic style of the message, the gender and age of the subjects and their familial relationship to the voice they were asked to identify.

Paper accepted (yes/no):

Paper number 32

Two Proposals

Jos Bouten and Tina Cambier Langeveld

1) A proficiency test on speaker verification. A 'Fake Case Proposal' was already proposed at the ENFSI Working Group for Forensic Speech and Audio Analysis meeting in Istanbul in 2003. We would like to extend an invitation to IAFPA conference attendees to join this test. The proposition will explain the goal and setup of this test.

Summary

Within the field of speaker identification, very little has been done on proficiency testing and validation. Some of the reasons for this are that:

(a) the field suffers from the use of widely varying methods, from acoustic-phonetic to semi-automatic to fully automatic, in different contexts (e.g. for evidential vs. investigative purposes).

(b) the expertise involved is to a large extent language-specific (in most methods), and the number of experts per country is limited.

Still, in light of the recent discussion on the value of other types of identification evidence (e.g. based on fingerprints, handwriting analyses, or earmarks), the need to examine the possibilities for validation of speaker identification is becoming increasingly clear, despite the − admittedly quite serious − complications that the above factors introduce. Therefore, we would like to conduct an evaluation of different methodologies and reporting strategies in an examination by different experts using the same set of materials. The aim of this evaluation will not be to evaluate the accuracy of several methods, since they are in many ways not directly comparable, but rather to document the different strategies that are used in different laboratories to come to a final report. A ‘fake case’ will be constructed in English, and experts will be asked to i) write a report as they normally would (i.e. as if this were a real case), and ii) provide additional information as to what led to the conclusions, or lack thereof, that were reached.

The actual construction of the fake case will be done by the NFI, which is also where all the reports and additional information will be gathered. We further propose that the reports be written in the native language of the expert(s) if this facilitates them. Reports that are not in English will be translated into English by official translating agencies, and then be sent back to the expert(s) who wrote them, so that they can assess the quality of the translation and perhaps correct any mistakes. The additional information describing the methodology of the speaker identification research can either be written directly in English or also in the native language of the expert(s).

2) A web accessible database called Magnet-O-Base.

Magnet-O-Base is a project that intends to build a web based database for storing and retrieving knowledge with respect to analogue tape authentication. This project was proposed in Istanbul at the 2003 meeting of the ENFSI Working Group for Forensic Speech and Audio Analysis. The presentation will explain the goal of the project and the costs involved.

Summary

A non-destructive method to inspect the magnetic layer of a tape which employs the Faraday effect (the ability of certain transparent materials to rotate the polarisation plane of the light passing through it in the presence of a magnetic field) is gaining popularity rapidly. Specifically the technique that uses so called ferri magnetic garnet films, also known as Analogovyj crystals and the MOA-KOV technique based on 'freezing' a magnetic fingerprint in a subtratum by flash heating it above its Curie temperature, makes it possible to take pictures showing the magnetic flux distribution on a tape in such detail that it enables the investigator to answer questions that couldn’t be addressed before. A knowledge tool based on a web accessible database (Magnet-O-Base) is proposed in which pictures of 'events' on magnetic tapes can be stored together with information related to the tape the event was found on and the recorder the event was produced with. The goal of this database is to serve as a reference for tape authentication investigations, for eductional purposes and for research and development purposes.

Jos Bouten is with the Netherlands Forensic Institute, Ministry of Justice, Rijswijk, The Netherlands.

Paper number 32

Tina Cambier Langeveld is with the Netherlands Forensic Institute, Ministry of Justice, Rijswijk, The Netherlands.