23rd CONFERENCE of the

INTERNATIONAL ASSOCIATION of FORENSIC PHONETICS and ACOUSTICS

Dear IAFPA 2014 delegates

We have the great pleasure of welcoming you to the 23rd conference of the *International Association of Forensic Phonetics and Acoustics* in Zürich/Switzerland!

This booklet introduces you to the local organisers and contains important information about the IAFPA venue, the programme and the abstracts. All in all, it should be everything you need for IAFPA 2014. If you wish to have to have more local information we recommend to use the interactive map on our webpage (link below).

IAFPA 2014 follows the traditional layout with three days of conference (Mon, Tue, Wed), finishing Wednesday around lunch time. We have a conference warm-up on Sunday (31st Aug) at 18:00 hrs and a conference banquet on Tuesday night. You will find all necessary information about this on the following pages.

We will not produce printed versions of this booklet since experience showed that delegates nowadays tend to read this information on their laptops or tablets. Should you wish any information of the booklet in print, we would like to ask you to print it prior to your arrival.

We sincerely hope that you will enjoy the conference as well as your stay in Zurich.

Volker Dellwo and the team of organisers at Zurich University

Also consult our webpage:

www.pholab.uzh.ch/iafpa2014.html

Who are we?

Any of the following people are happy to assist you at IAFPA 2014. Please do not hesitate to talk to us:

Stephan Schmid

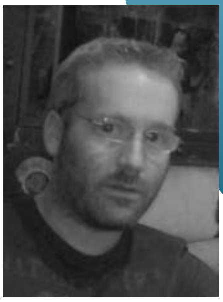
Adrian Leemann

Volker Dellwo
(chair)

Lei He

Marie-José Kolly

Ingrid Hove



Kostis Dimos

Sandra Schwab

Dario Brander

Daniel Friedrichs

IAFPA 2014 at a glance

	Sunday 31 Aug.	**Monday** 1 Sept.	**Tuesday** 2 Sept.	**Wednesday** 3 Sept.
DAYTIME		CONFERENCE START: 8:50 hrs END: 17:40 hrs VENUE I	CONFERENCE START: 9:00 hrs END: 16:00 hrs VENUE I IAFPA AGM* START: 16:00 hrs END:17:30 hrs VENUE I	CONFERENCE START: 9:00 hrs END: 12:50 hrs VENUE I
EVENING	Social Event START: 18:00 hrs END: open VENUE II	Committee meetings* START: 18:00 hrs END: open VENUE III	Social Event START: 19:00 hrs END: open VENUE III	

VENUE I:
Building RAI
Rämistrasse 74
8001 Zurich
(Rooms: J-31 and H-41)

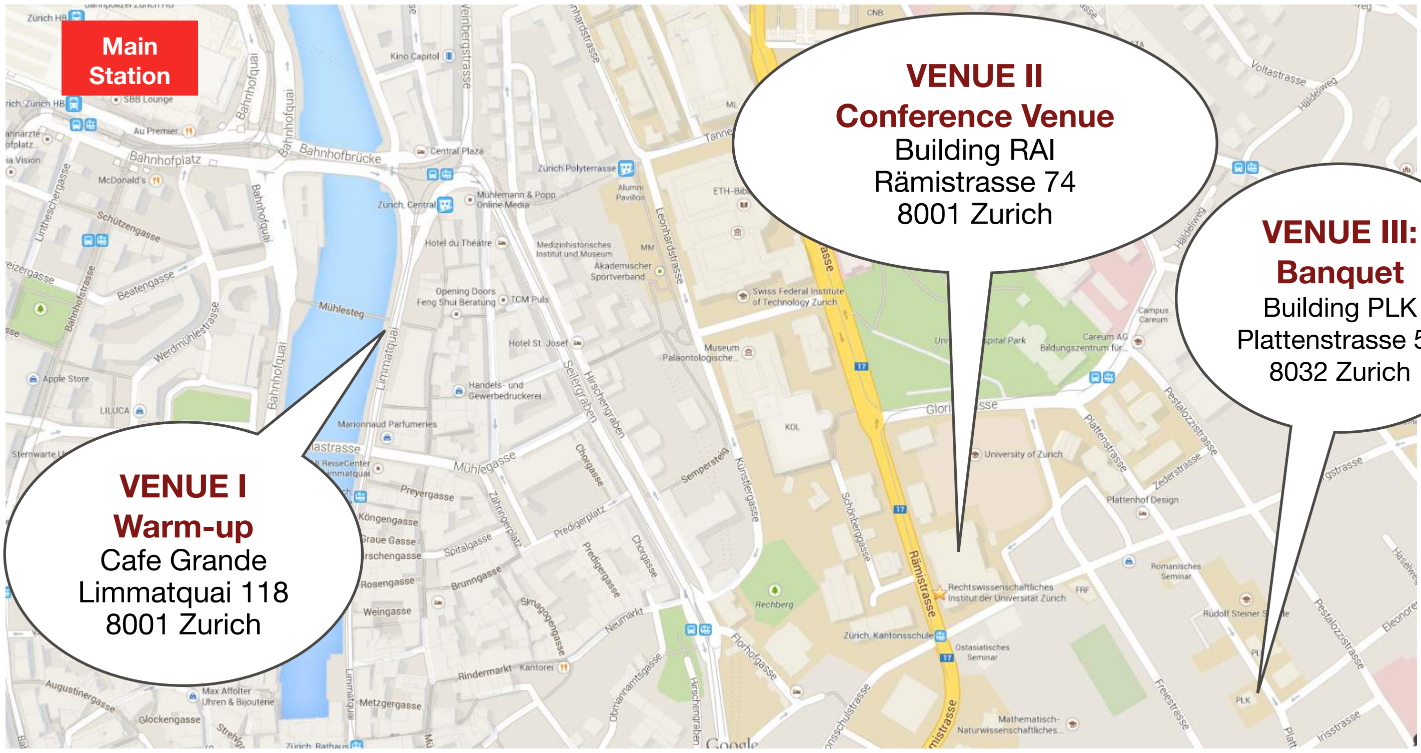
VENUE II:
Cafe Grande
Limmatquai 114
8001 Zurich

VENUE III:
Building PLK
Plattenstrasse 54
8032 Zurich

(see details on next pages)

* Please note that the IAFPA AGM is for IAFPA members only and 'Committee meetings' are for members of the Executive, Professional Conduct and Research Committees only

Map of the conference



The conference is centred around three places:

(a) **VENUE I:** The conference venue in building RAI of Zurich University at Rämistrasse 74 where the talks and poster sessions will take place.

(b) **VENUE II:** The conference warm-up in Cafe Grande at Limmatquai 118 on Sunday night.

(c) **VENUE III:** The conference banquet will be in building PLK on Tuesday night.

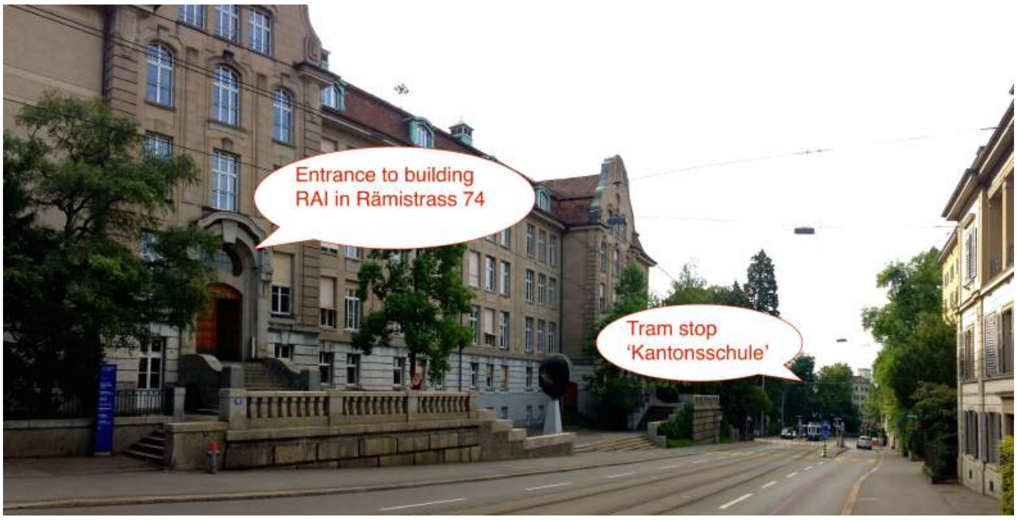
The area shown can be comfortably walked. The distance between the main station (top left) and building PLK (bottom right) takes 15 to 20 minutes by foot.

Scientific Events

VENUE I:

Building RAI of
Zurich University

Rämistrasse 74
8001 Zurich



All conference sessions (orals, posters) as well as the IAFPA AGM are in the building RAI of Zurich University in Rämistrasse 74. The nearest tram stop is '**Zurich Kantonsschule**'. The building can be reached from Zurich Main Station within 15 min by foot. When you reach the building make sure you take the left of the two rather similar entrance blocks (we will put up a sign).

ROOM 31 on FLOOR J

When inside the building, walk up to room 31 on floor J where you will find the registration desk (left). All coffee breaks and poster sessions will be held in this room as well.

LECTURE THEATRE 41 on FLOOR H

One floor below the registration room is the lecture theatre 41 where all oral sessions as well as the IAFPA AGM will be held.

Social Events

VENUE II: Sunday, 31 August 2014: Conference Warm-Up

We invite you to take part in the conference warm-up at **Cafe Grande on Limmatquai 118** (right) in the old town of Zurich.

18:00 hrs: We will meet in front of Cafe Grande for a short walk through the town to lake Zurich and back. Join us if you want to see some of Zurich's sights.

19:00 hrs: Cafe Grande will open for IAFPA and will serve a selection of drinks and light snacks which are included in the registration fee. Delegates looking for a proper meal might want to eat beforehand or afterwards in one of the many restaurants right round the corner from Cafe Grande in the old town. Cafe Grande will close at 10 pm.

VENUE III: Tuesday, 2 September 2014: Conference Banquet

We invite you from 19:00 hrs to the conference banquet at **building PLK of Zurich University in Plattenstrasse 54** (right). Building PLK is the home of the Department of Comparative Linguistics which also hosts some members of the Phonetics Laboratory. We are blessed in that we have a wonderful garden surrounding the building in which we will hold the banquet. The building is only 5 min walk away from the conference venue (see our interactive map on the IAFPA webpage under 'local information'). We will have a bbq with salads as well as some Ethiopian specialities (both meat and vegetarian). All drinks and food are included in the registration fee. The venue will be outside but we organised a heated tent in the garden in case it should rain. Nevertheless, we highly recommend to **bring some warm jumpers** as the evenings can sometimes get a bit chilly.

Registration

Registration takes place in building RAI (room 31 floor J, see above) and is open on Monday from 8:20 to 9:25 and then again during the coffee breaks and poster sessions. Upon registration you will receive:

- A conference name tag
- The wireless key
- A receipt of your payment
- A certificate of participation (upon request).

Please be reminded that payments in cash will need to be made either in CHF or in Euros. The following registration rates apply:

- **IAFPA members:**
  - regular: 200 CHF (170 Euros)
  - student: 100 CHF (85 Euros)
- **All others:**
  - regular: 240 CHF (200 Euros)
  - student: 120 CHF (100 Euros)

Additional guest-tickets for the conference dinner: 60 CHF (50 Euros)

We would appreciate if you had the respective amount ready for registration. As some participants may have not had a chance to obtain the registration fee in cash by Monday morning you can receive your name tag and internet key and make the payment by Tuesday at the latest.

Best student paper awards

By the end of the conference there will be a vote for the best student paper award. We will have one award for the best talk and another award for the best poster. To remember your favourite papers you can take notes in the following list of student contributions.

The winners of this prize will win a free registration for next year's IAFPA conference. Last year’s winner was Vincent Hughes from the University of York.

Your notes for TALKS		
Name	**Title**	**My rating**
ATKINSON, Nathan	Earwitness Identification: Is Just Once Enough?	
BRAUN, Almut	Feasibility of acoustic testing with fMRI for speaker recognition experiments	
ENZINGER, Ewald	A demonstration of the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case	
SANSEGUNDO, Eugenia	Forensic voice comparison using glottal parameters in twins and non-twin siblings	
VANKOVA, Jitka	Stability of short-term voice quality parameters in GSM	
WOOD, Sophie	Filled pauses as variables in speaker comparison: dynamic formant analysis and duration measurements improve performance	
WORMALD, Jessica BROWN, Georgina	Speaker profiling: An automatic method?	

Your notes for POSTERS

Name	Title	My rating
BAUMEISTER, Barbara	The influence of f0 on the perception of alcoholic intoxication	
BRANDER, Dario	Phonetic characteristics of hesitation vowels in Swiss German and their use for forensic phonetic speaker identification	
DIMOS, Kostis	An investigation of the rhythmic acoustic differences between normal and shouted voices	
ENZINGER, Ewald	Mismatch compensation in the evaluation of evidence under conditions reflecting those of an actual forensic-voice-comparison case	
FECHER, Natalie	Speaker discrimination based on ‘facewear speech’	
FEISER, Hanna	Perceptual voice similarity of related speakers: telephone and microphone recordings	
HE, Lei	Inter-speakers variability of intensity levels across syllables	
KOLLY, Marie-Jose	Speaker-individual rhythmic features in both L1 and L2 speech: implications for forensic voice comparison	
RENNING, Nancy	The Influence of Background Music on Perceived Seaker's Age	
SCHINDLER, Carola	Perceptual speaker discrimination based on German consonants	

Programme & Abstracts

The scientific programme of IAFPA 2014 will open on

Monday, 1st of September 2014
at 8:50

with some welcome notes by

the local organisers:
Volker Dellwo

the Zurich Institute for Forensic Sciences:
Peter Pfefferli
Thomas Ottiker

the International Association of Forensic Phonetics and Acoustics:
Peter French
Tina Cambier-Langeveld

Talks will start immediately thereafter (9:25)

	Monday	Tuesday	Wednesday
8:50	**Welcome notes** **Cambier & Vermeulen** *Gestalt: an undeniable part of human voice perception* (p. 13) **McDougall** *Listeners' perception of voice similarity in Standard Southern British English versus York English* (p. 63) **Vaňková et al.** *Stability of short-term voice quality parameters in GSM* (p. 75)	**Jessen** *Comparing MVKD and GMM-UBM applied to a corpus of segmented vowels in German* (p. 50) **Alexander et al.** *Zooplots for Speaker Recognition with Tall and Fat Animals* (p. 1) **Brown & Wornald** *Speaker profiling: An automatic method?* (p. 83) **Becker et al.** *Automatic Voice Comparison Performance in Forensic Casework* (p. 7)	**Gold & Hughes** *The correlation structure of speech parameters in Southern Standard British English* **Heeren et al.** *Exploring long-term formants in bilingual speakers* (p. 39) **San Segundo & Gomez-Vilda** *Forensic voice comparison using glottal parameters in twins and non-twin siblings* (p. 68) **Hove & Dellwo** *The effects of voice disguise on f0 and on formants* (p. 44)
10:40	*Coffee Break*	*Coffee Break*	*Coffee Break*
11:10	**Fraser** *Issues in the presentation of indistinct covert recordings a evidence in criminal trials* (p. 33) **Hirson et al.** *Easy straight upsay hospey' - the forensic decryption of Pig Latin* (p. 43) **Atkinson** *Earwitness Identification: Is Just Once Enough?* (p. 3) **Schreuder & Meyer** *Earwitness speaker identification and psychological responses* (p. 72)	**Watt et al.** *Ratings of 'threat' and 'intent' by listeners exposed to neutrally-worded utterances in five languages* (p. 79) **Dellwo & Brander** *Speaker and dialect effects in the dynamics of speech temporal characteristics* (p. 17) **Duckworth & McDougall** *Assessing the consistency of disfluency measures in characterising speakers* (p. 21) **Wood et al.** *Filled pauses as variables in speaker comparison: dynamic formant analysis and duration measurements improve performance* (p. 81)	**F.O.M.S, LINGUA** *Nativespeakerhood: A Subject Matter Revisited* (p. 62) **Kehrein & De Jong-Lendle** *„www.regionalsprache.de (REDE)" – a dialectological GIS for linguists and forensic phoneticians* (p. 52) **Leemann & Kolly** *Assessing the potential of crowdsourced 'Dialäkt Äpp' speech data for forensic phonetics* (p. 58) **de Jong-Lendle et al.** *Individual speaker characteristics of creaky phonation* (p. 15)
12:50	*Lunch*	*Lunch*	
14:00	*Poster Session I (see next page)*	*Poster Session II (see next page)*	
16:00	**Braun et al.** *Feasibility of acoustic testing with fMRI for speaker recognition experiments* (p. 11) **Gold & French** *An exercise in calculating numerical likelihood ratios and the practicalities of their implementation* (p. 36) **Forth & Akexander** *Content Comparison and Analysis (COCOA) of Contemporaneously Recorded Audio Material* (p. 31) **Enzinger & Morrison** *A demonstration of the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case* (p. 23)	***Annual General Meeting*** *(IAFPA Members only)* *19:00 Conference Dinner*	

Poster Session I	Poster Session II
Feiser & Draxler *Perceptual voice similarity of related speakers: telephone and microphone recordings* (p. 29) **Baumeister & Schiel** *The influence of f0 on the perception of alcoholic intoxication* (p. 5) **Dimos et al.** *An investigation of the rhythmic acoustic differences between normal and shouted voices* (p. 19) **Renning** *The Influence of Background Music on Perceived Speaker's Age* (p. 65) **Fecher & Watt** *Speaker discrimination based on 'facewear speech'* (p. 27) **Gómez et al.** *Dysphonic Voice Detection for Speakers' Biometry* (p. 38) **He & Dellwo** *Inter-speakers variability of intensity levels across syllables* (p. 41) **Leemann et al.** *Testing the effect of dialect imitation on suprasegmental temporal features* (p. 56) **Brander** *Phonetic characteristics of hesitation vowels in Swiss German and their use for forensic phonetic speaker identification.* (p. 9)	**Lindh et al.** *Effect of the Double-Filtering effect on Automatic Voice Comparison* (p. 60) **Enzinger** *Mismatch compensation in the evaluation of evidence under conditions reflecting those of an actual forensic-voice-comparison case* (p. 25) **Varošanec-Škarić et al.** *Comparison of similarity and dissimilarity indices between speech samples in filtered and non-filtered conditions for the speakers of the Croatian language* (p. 73) **Hughes et al.** *Modelling features for forensic speaker comparison* (p. 48) **Kolly et al.** *Speaker-individual rhythmic features in both L1 and L2 speech: implications for forensic voice comparison* (p. 54) **Rhodes** *Cognitive bias in forensic speech science* (p. 66) **Schindler et al.** *Perceptual speaker discrimination based on German consonants* (p. 70) **van der Vloed & Bouten** *NFI-FRITS: A forensic speaker recognition database* (p. 85) **Hove et al.** *Using the smartphone application 'Voice Äpp' to collect speech population data: implications for forensic phonetics* (p. 46)

Zooplots for Speaker Recognition with Tall and Fat Animals

Anil Alexander[1]*, Oscar Forth*[1]*, John Nash*[2]*, and Neil Yager*[3]
[1]*Oxford Wave Research Ltd,* [2]*University of York,* [3]*AICBT Ltd, United Kingdom*
`{anil|oscar@oxfordwaveresearch.com, neil@aicbt.com}`

Performance in speaker recognition is normally discussed using database-centric single figures of merit such as equal error rates. These metrics fail to capture the performances of individual speakers or speaker groups, which are very important in forensic speaker recognition. For instance, a recognition system that works well for male speakers may perform poorly for female speakers. Alternatively, a system may fail for speakers of a certain language or under a specific recording condition. The zoo-plot analysis, developed by Yager and Dunstone (2011), extends George Doddington's (1998) original classification of the biometric menagerie to categorise other difficult speakers. Under the original Doddington classification, *sheep*, who are 'normal' speakers and tend to match well against themselves and poorly against others, are the majority of the speakers within the database. *Goats* are speakers who are difficult to verify and tend to have low genuine match scores. *Lambs* generally match with high scores against other speakers and are thus easily impersonated, resulting in false accepts. *Wolves* easily impersonate other speakers, also resulting in false accepts. Yager and Dunstone extend this menagerie by taking both genuine and imposter performance into consideration, leading to four new 'animal' types: *chameleons*, *phantoms, doves* and *worms*. *Chameleons* always appear like others, receiving high scores for matches against themselves and others. *Phantoms* always receive low scores, so rarely match against themselves or others. *Doves* are the best possible users of a recognitions system, as they have high scores when matched against themselves and low scores when matched against others. *Worms* are the worst users of a biometric system, and are characterised by low genuine scores and high imposter scores.

Zooplot analysis is performed as follows: Select a group of speakers that represents a recording condition. From this set of speakers, select non-contemporaneous files for testing and training speakers. Ideally, there should be more than one file each for testing and training for the same speaker. For each speaker, match their training samples against all of their testing samples and compute their average genuine match score. Similarly, the mean of all the scores obtained by comparing his/her training samples with files from other speakers gives the average imposter score. In a two-dimensional quartile plot, as shown in Figure 2, the average genuine score is plotted against the average imposter score for all speakers. The users who fall within the four quartiles (top and bottom 25%) are assigned to the animal groups (*worms*, *chameleons*, *doves* and *phantoms*), with each set showing different characteristics.

In this work, we further extend the classification of these animals by characterising the speakers as 'tall/short' or 'fat/thin', depending on the variability of their genuine and imposter match scores (see Figure 3). For example, if a '*dove'* speaker has low genuine variability and high imposter variability, then he or she is a '*tall thin dove*'. Generally speaking, variability of match scores is symptomatic of an underlying problem, regardless of animal type. Therefore, the enhanced visualization adds a new dimension of independent and useful diagnostic information.

While single figures of merit like equal error rates provide information about performance of a system against a database as a whole, zooplot analysis can provide valuable insight into the properties of individual speakers and clusters of speakers in the database. It can help to identify potential algorithmic weaknesses of systems against certain classes of speakers, and can be used to adjust identification thresholds at an individual or group level. Preliminary research seems to suggest a link between certain aspects of voice quality and speaker categories in the zooplots. We recommend that zooplot analysis is done as speakers are added into a database, to help identify commonalities of speaker groups or algorithmic weaknesses of systems.

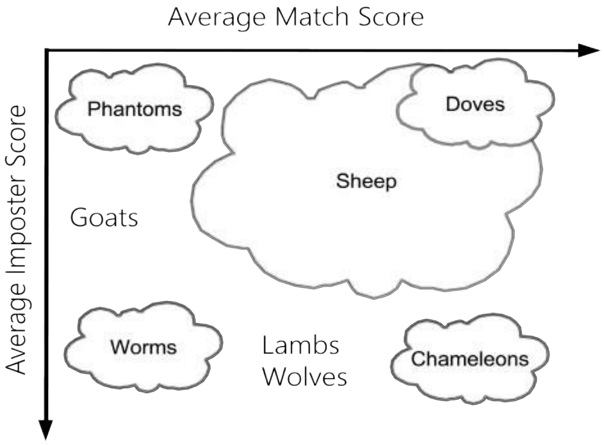


Figure 1: Illustration of a zooplots described in Yager and Dunstone 2011

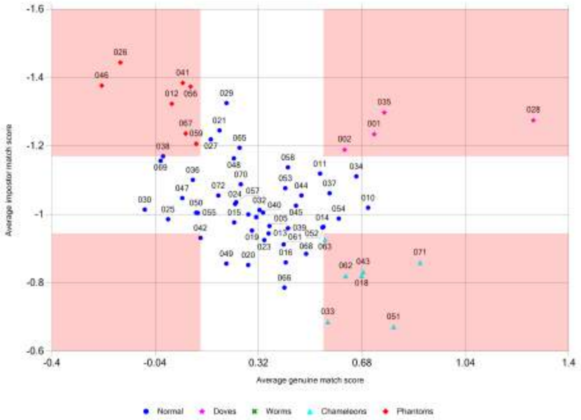


Figure 2: Zooplot using speakers from the IPSCO3 database using the VOCALISE spectral comparison

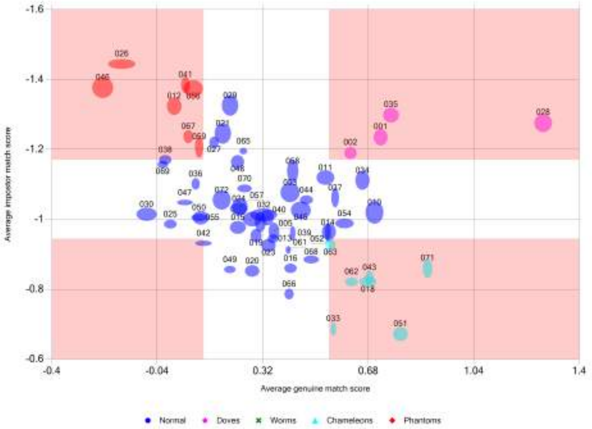


Figure 3: ‘Tall and Fat’ extension of Zooplot in Figure 2

References

N. Yager, and T. Dunstone, Biometric Systems for Data Analysis: Design, Evaluation, and Data Mining, 2009 Springer Press, ISBN-13:978-0-387-77625-5

G. Doddington, W. Liggett, A. Martin, M. Przybocki, D.Reynolds, Sheep, goats, lambs and wolves: a statistical analysis of speaker performance, Proceedings of IC-SLD’98, NIST 1998 Speaker Recognition Evaluation, Sydney, Australia, November 1998, pp. 1351–1354.

Eligible for student award

Earwitness Identification: Is Just Once Enough?

Nathan Atkinson

[1]*Department of Language and Linguistic Science, University of York, York, UK*
Nathan.atkinson@york.ac.uk

The validity of voice identification as a reliable process is to a great extent unknown. Research into factors affecting a listener's ability to identify an unfamiliar voice is ongoing and highlights a number of potential issues to consider in its forensic application (extensively reviewed by Broeders and Rietveld (1995). Where guidelines are in place governing the construction and delivery of a voice line-up, such as the MacFarlane Guidelines in the UK (Home Office Circular, 2003), the stipulation is that the earwitness should be presented with a selection of voices – suspect plus foils – and asked whether they believe any of those to belong to the criminal. This method appears uncontested and involves the earwitness selecting one voice on one occasion. The evidence provided by this method is binary, with one single selection made either in favour of the prosecution or defence.

A possible solution to this binary result is to test the reliability of the earwitness's identification by asking them to make more than one judgement. There are methodological and ethical problems with using either more than one line-up of foils or having a time delay between repeating identical tests. The present study will instead investigate the viability of short-term repeated tests.

Listeners will be exposed to one target voice and then hear a selection of six voices. Rather than hear the line-up of voices once, as in a traditional voice line-up, listeners will instead hear each voice three times in a different order without being told any voices are repeated. The utterances from any given speaker will differ so that the only link between the three samples is the voice. Each time a sample is heard, the listener will be asked to rate how likely they think it is that the voice belongs to the target speaker (0-10).

The target voice and each of the foils will thus be given three ratings per listener and so a listener's likelihood-of-being-the-target rating can be calculated for each voice (0-30). It is predicted that the target voice will produce a higher rating than any of the foil voices. Comparisons will be drawn with a control group, who will use a traditional single identification procedure. The rates of correct identifications will be compared, where a higher rating for the target voice than any of the foil voices is treated as indicative of a correct identification within the test group.

The ratio of the ratings given to the each voice compared to all others will be considered in order to assess whether the strength of these ratings provides an indication of voice identification reliability. It is predicted that higher ratios will be recorded for the target voice relative to the foil voices. The effect of repeated testing will also be considered, with ratios also calculated for each of the three phases in the test.

The results will be discussed and possible implications for earwitness identification will be considered.

References

Broders, A. & A Rietveld, A. (1995). Speaker identification by earwitnesses. *In:* J. P. Köster, J. P. & A. Braun (Eds.), *Studies in Forensic Phonetics,* Trier: Trier University Press.

Home Office Circular. (2003). *Advice On The Use Of Voice Identification Parades* [Online]. Available: https://www.gov.uk/government/publications/advice-on-the-use-of-voice-identification-parades.

Eligible for student award

The influence of f0 on the perception of alcoholic intoxication

B. Baumeister[1]*, F. Schiel*[1]

[1]*Institut für Phonetik und Sprachverarbeitung*
Ludwig-Maximilians-Universität, München, Germany

{bba|schiel}@phonetik.uni-muenchen.de

We report three perception tests concerning the ability of listeners to perceive alcoholic intoxication solely from the speech signal. Speech samples are taken from the German Alcohol Language Corpus (ALC)[1], a publicly available corpus with recordings of sober and intoxicated speech of 162 speakers. An earlier study (Baumeister et al., 2012) revealed that the majority of these speakers[2] (79.1%) increase their fundamental frequency (f0) while intoxicated. This study is concerned with the question whether f0 is also a relevant cue for the perception of intoxication.
We tested (1) the general ability of listeners to discriminate between sober and intoxicated stimuli pairs of the same speaker, (2) f0 compensated stimuli pairs to see if the discrimination rate decreases, and (3) sober speech stimuli with manipulated f0 to see if we can elicit the same effect as in real intoxicated speech.

Method and Results

All three tests are forced-choice discrimination tests where one pair of stimuli of the same speaker was presented at a time, and listeners were asked to pick the intoxicated stimulus. To compensate f0 effects in the stimuli for experiment (2), f0 of the intoxicated stimulus was adjusted in median f0 and range of f0 to the sober stimulus by up- or down-shifting and stretching or compressing the f0 contour. In the third experiment two sober stimuli of the same speaker were presented, but the f0 contour of one stimulus was up-shifted and stretched by 5%.
The mean discrimination rate of the basic discrimination test (1) is 61.8%, which is above chance. In a control group (two sober stimuli) listeners chose randomly as expected (49.2%). The mean discrimination rate in the compensation experiment (2) is 61.6% which - contrary to our expectations - does not differ significantly from (1). The average discrimination rate in experiment (3) is 52.5% and therefore slightly higher than chance ($p<0.1$).
The results suggest that f0 is not a relevant perceptual cue for listeners, although as shown in Baumeister et al. (2012) it seems to be a promising feature for the automatic detection of intoxication. Listeners seem to rely on other (maybe para-linguistic) features. Only if such other features are missing (as in experiment 3), a slight tendency to choose the stimulus with higher f0 can be observed. One possible explanation is that f0 is influenced by many other speaker states (such as stress, emotions) in a similar way as intoxication, and is therefore not reliable enough to reveal a speaker's intoxication.

References

Baumeister, B., Heinrich, C., Schiel, F. (2012). The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. *Journal of the Acoustical Society of America,* **132**, 442-451.

[1] For a detailed description of the ALC see Schiel et al. (2012)

[2] Only 148 speakers with a blood alcohol concentration higher than 0.05% were part of this study

Schiel, F., Heinrich, C., Barfüßer, S. (2012). Alcohol Language Corpus: The first public corpus of alcoholized German speech. *Language Resources and Evaluation* **46(3)**, 503-521.

Automatic Voice Comparison Performance in Forensic Casework

Timo Becker[1], *Gaëlle Jardine*[2], *Yosef Solewicz*[3] *and Stefan Gfrörer*[1]
[1]*Federal Criminal Police Office, Germany*
{timo.becker|stefan.gfroerer}@bka.bund.de
[2]gaelle.jardine@gmail.com
[3]*Israel National Police*
solewicz@police.gov.il

Introduction

One of the variables usually given to express the performance of a forensic voice comparison system is its error rates. A court that is presented with the results of ex perts' voice comparisons needs to be sure that what is stated as a system's theoretical error rate also applies to the one particular case being presented. This is not always the case: theoretical error rates are usually based on evaluations with speech corpora (the most common one being NIST (Przybocki et al. 2007)) that generally are of very much higher quality than forensic speech samples and therefore are likely to give better results.

Experts who therefore prefer to conduct their own evaluations in order to calculate error rates that they can reasonably claim to be appropriate to their given case are often confronted with the impossibility of collecting a big enough evaluation speech corpus, especially if their case contains channel mismatch.

Some experts will go ahead with the automatic voice comparison anyway and publish the "theoretical" error rate. They will obviously have to alert the court about the fact that in their specific case, at least if it contains typically forensic-quality speech samples, the actual error rate is unknown, but almost certain to be higher than the one being advertised. Other experts will prefer to avoid this uncertainty and choose not to use the system at all. In this case, however high-performing the system may be in theory, i.e. however low its theoretical error rate may be, in practice the system, not being used at all, has zero performance.

We would like to address this problem of real rather than theoretical performance and suggest a new way not only of defining performance already done by Bouten in 2012 (P.C. Jos Bouten), but of attaining better performance by combining two types of forensic voice comparison systems.

Combining two systems

Let us look at different examples of automatic voice comparison systems:

System **A** has a very low error rate, but this error rate is only known to apply to very specific recordings, let's say long telephone-quality, single-channel recordings.

System **B** has a higher error rate, but this error rate is known to apply to a much wider variety of recordings, let's say recordings that may be fairly short, noisy, and include channel mismatch between suspect and question files.

Following van Leeuwen & Brümmer (2007), we can assess the performance of sys tems **A** and **B** not in terms of their error rates, but in terms of actual information extracted, which we express in "bits". The interpretation of these bits is related to the common C_{llr} error measure which is the average information loss of a system. An average of 0 bits means a C_{llr} of 1 and vice versa. If we have a same-spea ker-com parison and the system outputs a likelihood ratio (LR) of infinity, one bit is extracted

(i.e. all the information available); if it outputs a LR of zero, zero bits are extracted (i.e. no information at all). For different-speaker-comparisons, the opposite is true.
Let's say System **A** only allows us to handle 10 cases a year, and for each it extracts 0.5 bits. This gives us a total extraction of 5 bits per year. Let's say System **B** only extracts 0.4 bits for each case, but it handles not just 10 but 25 cases a year. We obtain an average yearly total of 4 + 6 = 10 bits. What we can do now is set up a System **C** that combines Systems **A** and **B**: 10 cases will be extracted by its component **A**, 15 by its component **B**, and the total number of bits extracted yearly will be 5 + 6 = 11 bits. These examples are in line with real-life experience, as we will show using two current state-of-the-art voice comparison systems to simulate System **A** and a p-value approach which calculates scores without modelling intra-speaker variability (Solewicz et al. 2013) to simulate System **B**.

Discussion and Conclusion

While System **A**'s average performance remains better than **C**'s (and even more so of **B**'s), System **C**'s actual, total performance is better, since it handles two and half times as many cases as System **A** and is just as good as **A** in those cases that both can handle. However, depending on the field of interest, the expert might accept infor mation loss while processing a desired number of cases or vice versa.
What we would therefore like to suggest is using simpler voice comparison algo rithms, such as the one described in Solewicz et al. (2013), which produces scores without modelling intra-speaker variability. Such algorithms may at first seem like a step backwards when compared to state-of-the-art systems, but they could constitute a ma jor improvement to current practices in forensic speaker comparison, at least when com bined with these latter systems. For cases that meet certain, well-defined con ditions, the expert will be able to extract maximum information; for cases that don't, the expert will be able to extract less information, which is better than none at all.

References

Przybocki, Mark A., Martin, Alvin F. and Le, Audrey N. (2007): NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora – 2004, 2005, 2006. IEEE Transactions on Audio, Speech and Language Processing, **15**, 1951-1959.

Solewicz, Yosef A., Jardine, Gaëlle, Becker, Timo and Gfrörer, Stefan (2013): Estimated Intra-Speaker Variability Boundaries in Forensic Speaker Recognition Casework. Proceedings of Biometric Technologies in Forensic Science (BTFS) 2013, Nijmegen.

Van Leeuwen, David A. and Brümmer, Niko (2007): An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In Christian Müller (Ed.), *Speaker Classification I*, 330-353. Berlin. Springer.

Eligible for student award

Phonetic characteristics of hesitation vowels in Swiss German and their use for forensic phonetic speaker identification.

Dario Brander
Department of Comparative Linguistics, University of Zurich
dario_brander@gmx.ch

Hesitation vowels (commonly transcribed as *äh* / *ähm* in German-speaking regions) are one-syllable verbal utterances used to fill a speaking pause between two words or other linguistic elements. While the actual function of these sounds is still disputed (Clark & Fox Tree 2002, O'Connell & Kowal 2005), interest in using this type of verbal utterance for forensic phonetics has grown recently. An early study showed that interpersonal *and* interdialectal variations can be observed in f0 and F1/F2 of hesitation vowels (Alaoui in Jessen 2005, 273f.). Meanwhile, another study conducted under the instruction of the German Federal Police Office (Trouvain & Bauer 2005) confirmed these results and showed variability in the factors of usage (*äh* vs *ähm* and positioning in verbal utterances), temporal features (articulation rate hes/min) and intra-speaker variation. One further study (Klug & König 2012) additionally considered speakers' spread of data to be a speaker-specific factor as well.

The objectives of this contribution are to confirm the aforementioned results in a Swiss German speaker setting and to analyze further factors of potential inter-speaker-variation in hesitation vowels. The study features first results in the analysis of f0-ratio (the comparison of the initial and final 25% and central 50% of the vowel), F1-3 of the bilabial nasal in *ähm* and the temporal features general duration, vowel / nasal duration and vowel-nasal-ratio in *ähm.* Additionally, this study presents and discusses the possibility to optimize inter-speaker-variation by grouping the data material of speakers according to matters of usage (*äh* / *ähm* and positioning in a verbal utterance) and separate analysis under these conditions. In a small-scale comparative analysis, the data stability of one speaker's hesitation vowels is compared to the stability of [ə] of the same speaker in a read condition to determine if the analysis of hesitations alone yields clearer results than a general f0 and F1-3 analysis of a commonly used sound in Swiss German.

Methodologically, I proceeded as follows: 20 speakers of Zurich German (students, age 20-35) were recorded at the University of Zurich as part of Dellwo et al. 2012. The data was recorded in a sound treated booth. The participants partook in a 20-35 minutes long interview in which they were instructed to answer the questions freely. From this corpus, the 4 speakers of both gender groups with the highest number of hesitations were selected. Their hesitation vowels were extracted and analyzed. For the comparative analysis, a recording of a later phase of the recording sessions was chosen. One of the chosen speakers read 256 transcribed sentences of the former spontaneous recording session in Swiss German. His [ə]-vowels were extracted and their f0 and F1-3 analyzed. The standard deviations of those factors were then compared to the standard deviations of f0 and F1-3 of the hesitation vowels in the spontaneous condition.

References

Clark, Herbert H. / Fox Tree, Jean E. (2002): Using *uh* and *um* in spontaneous speaking. In: *Cognition, Vol. 84*. Amsterdam, Elsevier. 73-111.

Dellwo, Volker / Kolly, Marie-José / Leemann, Adrian (2012): Speaker identification based on temporal information: A forensic phonetic study of speech rhythm and timing in the Zurich variety of Swiss German. In: *Proceedings of the IAFPA 2012* (in print).
Abstract: http://p3.snf.ch/project-135287 [13.05.2013]

Jessen, Michael (2005): Conference Report: Annual Meeting of the International Association for Forensic Phonetics and Acoustics, Marrakesh, 3-6 August 2005. In: *Speech, Language and the Law 12 (2)*. Sheffield, Equinox. 279-280.

Klug, Katharina / König, Marie (2012): Untersuchung zur sprecherspezifischen Verwendung von Häsitationspartikeln anhand der Parameter Grundfrequenz und Vokalqualität. In: Hirschfeld, Ursula / Neuber, Baldur (Hg.): *Erforschung und Optimierung der Callcenterkommunikation*. Berlin, Frank & Timme. 175-194.

O'Connell, Daniel C. / Kowal, Sabine (2005): *Uh* and *uhm* Revisited: Are They Interjections for Signaling Delay? In: *Journal of Psycholingustic Research, Vol. 34, No. 6*. New York, Springer. 555-576.

Trouvain, Jürgen / Bauer, Dominik (2005): *Forschungsberichte zu den Projekten „Untersuchung der Sprecherspezifik der Grundfrequenz in Häsitationen und in äusserungsfinaler Position" und „Sprecherspezifik von gefüllten Pausen"*. BKA, March 2005.

Eligible for student award

Feasibility of acoustic testing with fMRI for speaker recognition experiments

Almut Braun[1], *Jens Sommer*[2], *and Andreas Jansen*[2]
[1]*Department of Phonetics, University of Marburg, Germany*
`almut.braun@staff.uni-marburg.de`
[2]*Section of BrainImaging, Department of Psychiatry and Psychotherapy, University of Marburg, Germany.*
`{jens.sommer|andreas.jansen}@med.uni-marburg.de`

Introduction:

The present study aims to investigate human speaker recognition ability while listeners are undergoing a functional MRI scan. Central questions are: To what extent - if at all - is it possible to do a more complex speaker recognition experiment within an MR scanner? If so, could different patterns of BOLD (blood-oxygen-level-dependent) activation and deactivation be linked to a listeners‘ performance in a speaker recognition task? Do familiar voices evoke BOLD activations in other areas of the brain than voices which have just been heard one time before?

In previous studies, voice coding has been associated with the superior temporal sulci (STS) and the inferior frontal cortex (Andics et al. 2013), and voice recognition was associated with the middle and posterior STS, the right ventrolateral prefrontal regions and the insular cortex, the anterior temporal pole (Andics et al. 2010).

This is a feasibility study. It will be tested whether it is generally possible to do a speaker recognition experiment within the noisy environment of a 3-tesla MR scanner. Different settings of the scanner as well as different types of headphones (electroacoustic/pneumatic) have been tested to reduce the subjective noise level. It was reported that the best noise reduction could be obtained when the listener was wearing electroacoustic headphones (mr-confon). Additionally, the acoustic condition was improved by wrapping the participant’s head with special foam material inside the head coil. Further improvements could be achieved by separating the noise and voice frequencies by adjusting the scanning parameters (e.g. echo time, repetition time, field of view, matrix)
If the feasibility study reveals no weaknesses in the local setup, a follow-up study with blind and sighted listeners will be carried out. Gougoux et al. 2009 found different activation patterns for blind and sighted listeners in a voice discrimination task (same/different speaker).

Method:

The experiment consists of two parts. In the first part, a sound file with 15 spontaneous voice samples of different male speakers is played to the listeners while lying inside the MR scanner. Ten of these voice samples come from famous speakers which are supposed to be recognized easily, five samples are voices which the listeners have never heard before. Each voice sample (duration: 30 seconds) is followed by a silent interval of 10 seconds to ensure that a baseline can be established. The famous speakers are selected according to a prior experiment with other participants. Famous striking voices are included in the experiment to provoke extreme reactions to estimate a general effect size for the given task.
In a second part, the participants undergo a second fMRI scan. This time, they have to listen to another sound file. This sound file consists of voice samples of 3 non-famous speakers they had heard in the first part of the experiment and 6 new (unknown) voices. Participants are

asked to indicate which of the voices they have heard before.

References

Andics, A., McQueen, J.M., Petersson, K.M.,Gal, V., Rudas, G., and Vidnyanszky, Z. (2010). Neural mechanisms for voice recognition.*Neuroimage* **52**, 1528–1540.

Andics, A. , McQueen,J.M., Petersson, K.M.. (2013) Mean-based neural coding of voices. *Neuroimage* **79**, 351-360

Gougoux, F.; Belin, P.; Voss, P.; Lepore, F.; Lassonde, M.; Zatorre, R.J. (2009): Voice perception in blind persons: a functional magnetic resonance imaging study. *Neuropsychologia* **47** (13), 2967-2974.

Gestalt: an undeniable part of human voice perception

Tina Cambier-Langeveld[1] *and Jos Vermeulen*[2]

[1]*Ministry of Security and Justice, Immigration and Naturalisation Service*
GM.Cambier.Langeveld@ind.minvenj.nl

[2]*Ministry of Security and Justice, Netherlands Forensic Institute*
j.vermeulen@nfi.minvenj.nl

The discriminating power of voice quality vs. voice quality in FSC

Voice quality is described by Abercrombie (1967:91) as "those characteristics which are present more or less all the time that a person is talking: It is a quasi-permanent quality running through all the sound that issues from his mouth." In a broad sense, voice quality is the total product of laryngeal phonation and supralaryngeal filtering, radiated from the mouth and nose and resonating through the soft tissue, bony structures and cavities in chest, neck and head. Given that humans can identify individuals by their voice alone, the discriminating power of whatever it is that we perceive as 'voice' must be quite good. The value of voice quality for FSR is generally recognised (Hollien 1990, Baldwin and French 1990).

From this viewpoint, it is remarkable that the description of voice quality generally receives little attention in expert reports on forensic speaker comparisons (FSC). In this paper, we will first present the role of voice quality in the reports collected by Cambier-Langeveld (2007). We compare this with the review by Nolan (2005) of approximately 30 cases in the British Isles. Nolan found that comments by forensic phoneticians on voice quality tend to be limited to observations like 'there were similarities in voice quality'. The expert reports contained only occasional evidence of componential analysis of voice quality.

The challenge and a solution

In our presentation, we will argue that a featural protocol for assessing voice quality, such as the Laver Vocal Profile Analysis scheme (Laver 1980), cannot capture the uniqueness of a voice, simply because the voice is typically processed by human listeners holistically, i.e. as Gestalt (Kreiman and Sidtis 2011).

We view Gestalt processing as an inherent part of auditory perception that cannot be 'switched off' at will; it is a real and important phenomenon in speaker recognition. We argue that a report based only on componential analysis does not really do justice to the perceptual mechanisms that are at work.

Central to this paper is the challenge to give Gestalt perception a place in FSC. To meet this challenge, we re-address the so-called 'blind grouping' method. This method has been presented to IAFPA earlier as a means to fight confirmation bias (Cambier-Langeveld and van der Torre 2004, Schreuder 2011). This method might also be an answer to the call for testing the expert's performance under conditions reflecting those of the case under investigation (Morrison, in press).

Blind grouping does not require verbal-analytic terminology, but requires the expert to compare anonymised fragments and arrange them into groups based on same-speaker and different-speaker judgements. It allows the forensic expert to use any strategy to reach a result, including pattern recognition and feature analysis. This method is proposed as a supplement to other methods. The presentation will include a demonstration.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Baldwin, J. and P. French (1990). *Forensic Phonetics*. London and New York: Pinter.

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification: Results of a collaborative exercise. *The International Journal of Speech, Language and the Law* 14(2), 223-243.

Cambier-Langeveld, T. and E.J. van der Torre (2004). Fighting the Confirmation Bias: Blind Grouping. *Presentation at IAFPA 13th Annual Conference*. Helsinki, Finland, 28-31 July.

Hollien, H. (1990). *The Acoustics of Crime. The New Science of Forensic Phonetics*. New York and London: Plenum Press.

Kreiman, J. and D. Sidtis (2011). *Foundations of Voice Studies. An interdisciplinary approach to voice production and perception*. Chichester: Wiley-Blackwell.

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

Morrison, G.S. (in press). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. Science and Justice, http://dx.doi.org/10.1016/j.scijus.2013.07.004

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

Schreuder, M. (2011). Expectancy bias and forensic speaker identification. *Presentation at IAFPA 20th Annual Conference*. Vienna, Austria, 24-28 July.

Individual speaker characteristics of creaky phonation: a comparative study of English and Urdu

Gea de Jong-Lendle[1], *Sarmad Hussain* [2], *and Francis Nolan*[3]
[1]*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg,Germany*
gea.dejong@staff.uni-marburg.de
[2]*CLE, Al-Khawarizmi Institute of Computer Science, UET, Lahore, Pakistan*
sarmad.hussain@kics.edu.pk
[3]*Institu Dept. of Theoretical & Applied Linguistics, University of Cambridge, UK*
fjn1@cam.ac.uk

The definition and the role of creak

Creak, also called vocal fry or glottal fry, has been defined by Catford (1964:32) as: ‘ Low frequency (down to about 40Hz) periodic vibration of a small section of the vocal folds.‘ Although the phenomenon of creak has received quite some attention in the past (Michel and Hollien 1968, Hollien et al 1966, Moore 1971, Coleman 1963) the exact physiological mechanics of creak are still unknown. The typical auditory quality of creak can be described as a series of separate taps in rapid sequence and it is therefore easily audible. In a spectrogram creak shows up as slightly irregular vertical striations. See Laver (1980) for a detailed overview on creak. In many tone languages syllables with low or falling tones are often accompanied with this type of phonation. In languages like English, creak has often been associated with the paralinguistic feature of turn taking; a speaker can use a falling intonation and creak as a signal that his/her turn has been accomplished, yielding the floor to the listener. When a speaker uses creak throughout an entire utterance or sentence, it is assumed that he/she is bored (Laver 1980:126) or wants to signal indifference.

Creak in a forensic context

Despite the fact that creak can almost be considered as a normal part of communication and is therefore present in a large number of speech recordings, it has received comparatively little attention within the forensic community. In Hudson et al. (2007) an early attempt was made as part of a larger project on the F0 statistics of 100 Standard Southern British English (=SSBE) speakers. Here it was found that, with the minimum threshold of 50Hz, creak showed up in the histogram in 66% of the cases. In 34% of the cases it can be assumed that either 1) creak is absent, or 2) creak was not measured as it was below the lower F0 threshold. The latter possibility being the more likely one of the two. In 3% of the cases the F0-mode was actually found in the creaky part of the H0 histogram instead of the F0- range covering the modal voice. In other words, in the bimodul distribution the creak-peak was higher than the modal peak. Observing the different bimodul histogram patterns, all speakers could be largely divided into 5 different groups.

Objectives

The aim of this investigation is to explore into more detail the speaker identifying potential of creak and creaky phonation.

The following questions are of interest:

1. Per speaker the percentage of creak in a 3 min. segment of normal spontaneous speech

2. F0 histograms when F0 measured with a F0 minimum threshold below 50Hz
3. Is the starting F0-point for creak the same for a falling intonation pattern as for a rising pattern?
4. Could the above patterns for the use of creak be different across languages/ethnological backgrounds? For example, what happens when speakers of a particular language speak at a much higher average speaking F0? Can one expect to find less creak? Here, we would like to compare English with Urdu as it was found that in SSB-English the average F0 mode was 102Hz (Hudson et al. 2007) and in Lahore-Urdu this value was calculated to be 129Hz (deJong et al. 2012).

As creak is difficult to measure due to its irregular behaviour, a 'quality control' is included to compare auditory detection of creak with f0 values produced by PRAAT. In addition, stretches of low f0 are examined to check that these correspond to auditory judgments, and aren't just e.g. halving errors occurring for other reasons.

If the speaker identifying potential is confirmed measuring groups of a small size (n=30), data could then be generated for a larger population. The forensic community could then be provided with probabilistic data concerning the use of creak.

Materials

Speakers: 30 male speakers of SSB-English + 30 speakers of Lahore-Urdu
Speaking style: spontaneous
Software: PRAAT (histograms are produced by a PRAAT script)

The SSB-English speaker recordings come from the DyViS -project in Cambridge. For a detailed description see Nolan et al. 2006a. The Lahore-Urdu speaker recordings come from the URDU-project in Lahore. For a detailed description see Sarfraz et al. 2010.

References

Catford, J.C. (1964) Phonation types: the classification of some laryngeal components of speech production, in Abercrombie, D. et al (19064) pp. 26-37.

Coleman, R.F. (1963) Decay characteristics of vocal fry, *Folia Phoniatrica* **15**: 256-63.

deJong,G., Hussain, S., Irtza, S. and Hudson, T. (2012) F0 statistics for speakers of Urdu, Proceedings, IAFPA 2012, Santander, Spain.

Hollien, H., Moore, P., Wendahl, R.W. and Michel, J.F. (1966) On the nature of vocal fry, *Journal of Speech and Hearing Research*, **9**:245-7.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. and F. Nolan, (2007). F0 statistics for 100 young male speakers of Standard Southern British English. ICPHS conference proceeding, In: J. Trouvain and W. Barry.

Laver, J. (1980) *The phonetic description of voice quality*, Cambridge University Press.

Michel, J.F. and Hollien, H. (1968) Perceptual differentiation in vocal fry and harshness, *Journal of Speech and Hearing Research*, **11**:439-43.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2006a). ‘A forensic phonetic study of ‘dynamic’ sources of variability in speech: the DyViS project.’ In: P. Warren and C.I. Watson (eds.), Proceedings of the 11th Australasian International Conference on Speech Science and Technology, 6-8 December 2006, Auckland: Australasian Speech Science and Technology Association, 13-18.

Sarfraz, H., Hussain, S., Bokhair, R., Raza, A.A., Sarfraz, Z., Pervez, S., Mustafa, A., Javed, I. and Parveen, R. (2010). Speech corpus development for a speaker independent spontaneous Urdu Speech recognition system, *Proceedings of O-COCOSDA*, Kathmandu, Nepal.

Speaker and dialect effects in the dynamics of speech temporal characteristics

Volker Dellwo & Dario Brander
Phonetics Laboratory, Department of Comparative Phonetics, University of Zurich
volker.dellwo@uzh.ch

It has been repeatedly demonstrated that speakers vary in their speech rhythmic characteristics and that such characteristics might be cues to the identity of a speaker and as such relevant to forensic speaker identification (Leeman et al., 2014, Dellwo et al., 2012). A shortfall with measures of speech rhythm so far is that they are based on a durational characterization of speech intervals (e.g. a syllable, a vocalic or a consonantal interval) that is averaged over the entire utterance. For example, typical measures of speech rhythm are based on standard deviations of speech intervals (e.g. the standard deviation of consonantal intervals, Ramus et al., 1999) or the average differences between syllables in a phrase (e.g. the Pairwise Variability Index, Grabe and Low, 2002). This does not take into account the dynamics with which speakers might vary temporal characteristics of speech over the course of an utterance.

To test whether there is reason to believe that inter-speaker variability exists in the dynamics of syllable durations within an utterance we analyzed the syllable durations in 256 sentences produced by 4 male speakers of Swiss German (64 sentences each) from two dialect regions (2 Bern, 2 Zurich). The sentences were a Swiss version of the Coordinate Response Measure Corpus (Moor, 1981, Bolia, 2000) recorded in our lab in Zurich, which means that all speakers uttered structurally identical sentences of the exact same number of syllables (16) that only varied in the choice of some lexical items. To calculate the syllable duration dynamics between speakers we first calculated a proportional duration for each syllable (duration of a syllable in percent re the total duration of the utterance) and then calculated the difference in duration between consecutive syllable pairs (15 pairs); henceforth: 'Proportional syllable differences (in %)'. Figure 1 contains the mean of the proportional differences for each speaker (red = Bern, blue = Zurich; values averaged over 64 productions per speaker). A value around 0 indicates that the syllable pair was produced with about equal duration for each syllable, a positive value indicates that the first syllable in a pair was longer than the second, a negative value that the first syllable was shorter than the second.

Results revealed: (a) The largest differences were obtainable in the first part of the phrase up to syllable pair 13. This means that the phrase final part (i.e. phrase final lengthening) did not vary between speakers nor between dialects. (b) There were possible speaker and dialect effects in different parts of the sentence: Between pair 1 and 9 the differences varied strongly between speakers irrespective of their dialect. Between pair 10 and 13 the differences showed some similarities as a function of dialect.

One of the main shortfalls of this study is that it relies on highly controlled material (speakers uttered sentences of the exact same structure) and that it is based on syllable durations, a rather ambiguous durational interval in speech outside the laboratory. We are now working on methods to compare speaker specific aspects of temporal dynamics between sentences of a different structure and using different temporal intervals.

References

Dellwo, V., Leemann, A. und Kolly, M.J. (2012) Speaker idiosyncratic rhythmic features in the speech signal. In: Electronic Proceedings of Interspeech, Portland/Oregon/USA.

E. Grabe, E.L. Low (2002) Durational variability in speech and the Rhythm Class Hypothesis C. Gussenhoven, N. Warner (Eds.), Laboratory Phonology, vol. 7Mouton de Gruyter, Berlin/New York, 515‐ 545.

Leemann, A., Kolly, M.-J., Dellwo, V. (2014) Speaker-individuality in supra-segmental temporal features: Implications for forensic voice comparison. In: Forensic Science International (238), 59-67.

Moore, T. J. (1981) Voice communication jamming research. In: AGARD, Conference Proceedings 311: Aural Communication in Aviation (AGARD), Neuilly-Sur-Seine, France, pp. 2:1–2:6.

F. Ramus, M. Nespor, J. Mehler (1999) Correlates of linguistic rhythm in the speech signal. In: Cognition (73), 265-292.

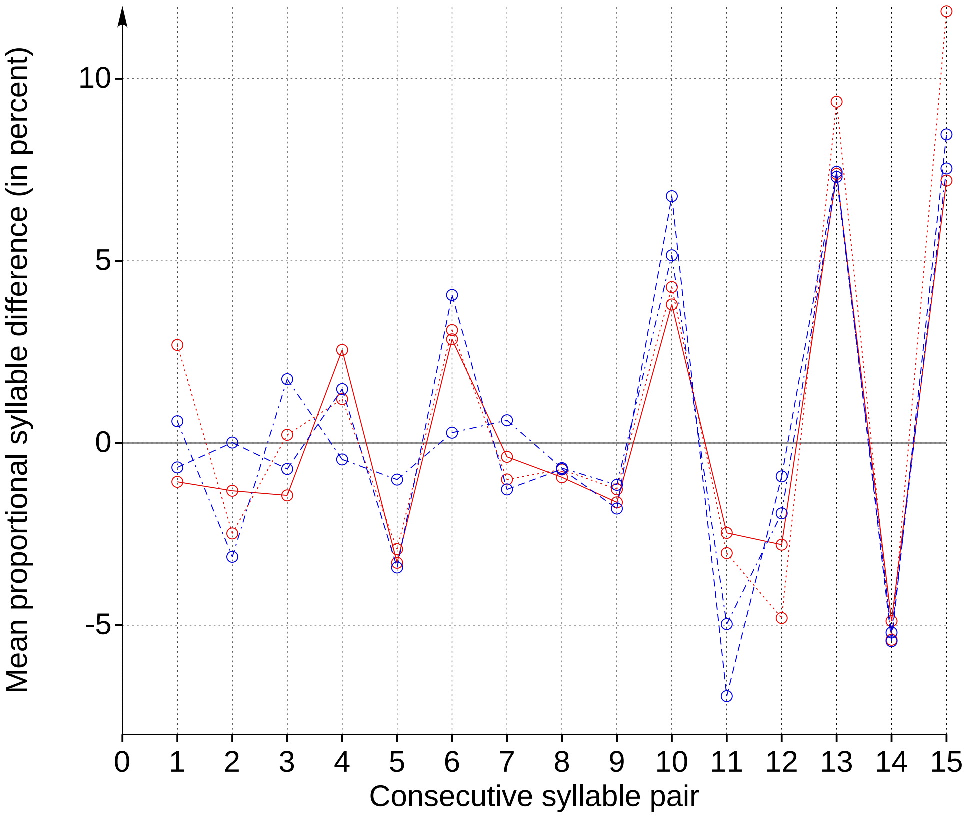


Figure 1: Graph showing the mean proportional syllable difference (in %) for each consecutive syllable pair from the first (1) to the last syllable pair (15) of the 16 syllable sentences.

Eligible for student award

An investigation of the rhythmic acoustic differences between normal and shouted voices

Kostis Dimos, Lei He, Volker Dellwo,
Phonetics Laboratory, University of Zurich, Zurich, Switzerland
{kostis.dimos|lei.he|volker.dellwo}@uzh.ch

This study aims at investigating rhythmic characteristics of shouting and comparing them to normal speech. There have been a number of studies examining segmental characteristics (Traunmüller & Eriksson 2000) in high vocal effort speech which revealed considerable differences between the two conditions; however there has been little research in prosodic, and especially, rhythmic characteristics of shouted speech. We expect that shouted speech will exhibit distinctive rhythmic characteristics as the control over the articulators varies a lot.

Ten, gender balanced, Zurich German speakers have been recorded producing semi-spontaneous utterances in both normal and shouted modalities. By semi-spontaneous, we mean that the researchers had control over the sentence structure of the utterances while at the same time keeping the data ecologically valid (Post & Nolan 2012). Following the methods described in Kainada and Baltazani (2013), we have created 15 pictures that the participants will have to describe. The material was divided in three equal groups, each containing 15 utterances, depending on the lengths of the utterances. The structure of the sentence was controlled by instructing the participants to always name the subjects displayed in the picture.

Well-established rhythm metrics (ΔC, %V, rPVI-C, nPVI-V, VarcoC, VarcoV, deltaPeak, VarcoPeak. Ramus et al., 1999, Grabe & Low, 2002, Dellwo, 2006, Dellwo et al., 2012a) were employed to measure the temporal characteristics of the shouted voice and the normal voice. Manual segmentation and labelling have been carried out by the research assistant together with the investigators based on the criterion described in Dellwo et al. (2012a, 2012b, 2012c).

Our preliminary findings indicate an effect on speech rhythm regularity. Additionally, between-speakers variability appears to be significant in shouted voice, as we have found in our previous research on normal voice (Dellwo et al. 2012a, Leemann et al. 2014).

References

Dellwo, V. (2006). Rhythm and speech rate: a variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), *Language and Language Processing*, 231-241, Frankfurt: Peter Lang.

Dellwo, V., Leemann, A. and Kolly, M.-J. (2012a). Speaker idiosyncratic rhythm features in the speech signal. In *Interspeech*, Portland, USA.

Dellwo, V., Kolly, M.-J. & Leemann, A. (2012b). Speaker identification based on speech temporal information: A forensic phonetic study of speech rhythm in the Zurich variety of Swiss German. Abstract presented at IAFPA 2012, Santander/Spain.

Dellwo, V., Schmid, S., Leemann, A., Kolly, M.-J. & Müller, M. (2012c). Speaker identification based on speech rhythm: the case of bilinguals. Abstract presented at PoRT2012, Glasgow/UK.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and rhythm class hypothesis. In N. Warner & C. Gussenhoven (Eds.), *Papers in Laboratory*

Phonology 7, 515-543, Berlin and New York: Mouton de Gruyter.

Kainada, E. and Baltazani, M. (2013). Evaluating methods for eliciting dialectal speech. In M. Janse, B. Joseph, A. Ralli and M. Bagriacik (Eds.), *Proceedings of the 5th International Conference on Modern Greek dialects and Linguistic Theory*, 101-123.

Leemann, A., Kolly, M.-J., and Dellwo, V. (2014). Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Sci. Int.*, **238**, 59-67.

Post, B. and Nolan, F. (2012). Data collection for prosodic analysis of continuous speech and dialectal vatiation. In Cohn, A. G., Fougeron, C. and Huffman, M. K. (Eds.), *The Oxford Handbook of Laboratory Phonology*, 538-547. Oxford: Oxford University Press.

Ramus, F., Nespor, M. and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **73**, 265-292.

Traunmüller, H., and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust Soc Am.*, **107**, 3438.

Assessing the consistency of disfluency measures in characterising speakers

Martin Duckworth[1] *and Kirsty McDougall*[2]

[1]*Duckworth Consultancy Ltd, UK*
msd@duckworth-consultancy.co.uk

[2]*University of Cambridge, UK.*
kem37@cam.ac.uk

This paper presents further results from an ongoing programme of research investigation the potential use of disfluency measures in forensic speaker comparison.[1] At IAFPA 2012 and 2013 results of an investigation of individual differences in disfluency behaviour in the speech of 20 male speakers of Standard Southern British English from the DyViS database was presented. Disfluency features analysed included filled and silent pauses, repetitions, prolongations and self-imposed speech interruptions. Although the overarching hypothesis behind this work is that disfluencies might have a speaker-specific aspect to them, it is acknowledged that disfluency events are also related to other cognitive and behavioural phenomena such as speech planning, conversational management and prosody. Therefore disfluency, rather like fundamental frequency and speaking rate may be affected by the content and context of speech.

Speaker-specific patterns were observed in the types of disfluency features used and how often they used them. These patterns showed a degree of within-speaker consistency across the two speaking styles examined: a mock police interview and a telephone call with a friend. This suggests that, despite occurring in different contexts, the amount and type of disfluency behaviour may be relatively consistent within a given speaker.

While disfluency features appear to offer an additional source of individual information about a speaker, the degree of subjective judgement involved in their identification and categorisation may undermine the usefulness of this analysis. In the study described above, the disfluency features were transcribed and categorised by a single analyst. For the present study, a subset of the data (5 speakers, interview style) is reanalysed by two additional analysts and the results of the three analysts compared in order to evaluate the consistency of disfluency feature measurements across analysts.

The two new analysts undertook training with the first analyst to become familiar with the criteria for identifying each disfluency feature type and the system for coding them. At a subsequent meeting, the analysts discussed their experiences of using the categorisation system and jointly decided on revised criteria for the identification of features which had proved ambiguous or problematic. Each analyst then worked independently on refining his or her own coding record using the improved categorisation criteria. The three final sets of disfluency measurements will be compared to assess the inter-analyst consistency of the method and implications of the findings for forensic casework will be discussed.

Preliminary results comparing measurements made by two of the analysts indicate high levels of inter-analyst correspondence for filled pause categories, silent pause categories, repetition categories and self-interruption categories. Some other categories were problematic however and we surmise that they may be less perceptually salient than others and/or pose a particular cognitive load in their identification. We will discuss how features may be defined in order to improve the consistency with which they may be identified.

References

M. Duckworth and K. McDougall (2012) ‘Developing disfluency profiles for individual speakers: a study of Standard Southern British English.’ Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Santander, 5-8 August 2012.

M. Duckworth and K. McDougall (2013) ‘Individual Differences in Fluency Disruptions: A Cross-Style Investigation.’ Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Tampa, Florida, 21-24 July 2013.

F. Nolan, K. McDougall, G. de Jong & T. Hudson (2009) The *DyViS* Database: Style-Controlled Recordings of 100 Homogeneous Speakers for Forensic Phonetic Research. *International Journal of Speech, Language and the Law,* **16.1**, 31–57.

[1] ‘The speaker-specificity of fluency disruptions’ was supported by an IAFPA research grant.

Eligible for student award

A demonstration of the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case

Ewald Enzinger[1,2] *and Geoffrey Stewart Morrison*[1,3]
[1]*School of Elec. Eng. & Telecom., University of New South Wales, Sydney, Australia*
[2]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*
[3]*Department of Linguistics, University of Alberta, Edmonton, Canada*
ewald.enzinger@oeaw.ac.at, geoff.morrison@forensic-evaluation.net

This presentation demonstrates the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case. This includes consideration of the relevant prosecution and defence hypotheses to address in this case, selection of data reflecting the adopted defence hypothesis, simulation of recording conditions reflecting those of the suspect and offender recordings in the case, quantitative measurement and statistical modelling to calculate a likelihood ratio given the relevant hypotheses and under recording conditions reflecting those of the case, and empirical testing of the validity and reliability of the resulting system given the relevant hypotheses and under recording conditions reflecting those of the case. As such, this provides a practical demonstration of a forensic voice comparison conducted under a paradigm which we have previously espoused (see Morrison, 2013, and Morrison & Stoel, 2013, for recent summaries of the paradigm).

There was no dispute in this case that the suspect and the speaker on the offender recording were adult male Australian English speakers, and we were able to draw samples from a database of multiple non-contemporaneous recordings of adult male Australian English speakers. The database included high-quality recordings of speech from an information-exchange-via-telephone task and a face-to-face interview task, which best reflected the speaking styles in the offender and suspect recordings respectively. We will discuss how the defence hypothesis in this case was further refined from adult-male Australian English speaker and how a relevant subset of the database was selected.

The offender recording in this case was of a landline telephone call made to a call centre. As well as telephone transmission, it included background noise at the call centre, and it was saved in a compressed format. The suspect recording was of a police interview conducted in a reverberant room with ventilation noise and saved in a compressed format. The presentation will include a description of how we simulated these conditions so that database recordings could be converted and used to train and test statistical models under conditions reflecting those of the case. We will play audio recordings which illustrate the steps in simulating the recording conditions.

The presentation will also include brief descriptions of: the procedures used to make quantitative measurements of the acoustic properties of the voices on the recordings, statistical modelling procedures used to calculate likelihood ratios (details of channel compensation techniques is the subject of another proposed presentation), and the procedure used to empirically test the validity and the reliability of the system. Finally, the results of system testing will be presented.

References

Morrison, G. S., & Stoel, R. D. (2013 online). Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come? Australian Journal of Forensic Sciences. doi:10.1080/00450618.2013.833648

Morrison, G. S. (2013 online). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. Science & Justice. doi:10.1016/j.scijus.2013.07.004

Eligible for student award

Mismatch compensation in the evaluation of evidence under conditions reflecting those of an actual forensic-voice-comparison case

Ewald Enzinger[1,2]
[1]*School of Elec. Eng. & Telecom., University of New South Wales, Sydney, Australia*
[2]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*
ewald.enzinger@oeaw.ac.at

This presentation demonstrates the application of mismatch compensation techniques in the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case. Several approaches developed in automatic speaker recognition research are considered for use in a forensic-voice-comparison analysis to reduce variability in quantitative measurements made of the acoustic properties of the voices on the suspect and offender recordings caused by mismatched recording conditions. Other aspects of the forensic analysis such as the consideration of the relevant prosecution and defence hypotheses to address in this case, selection of data reflecting the adopted defence hypothesis, statistical modelling, and likelihood ratio calculation are the subject of another proposed presentation.

The offender recording in this case was of a landline telephone call made to a call centre. As well as telephone transmission, it included background noise at the call centre, and it was saved in a compressed format. The suspect recording was of a police interview conducted in a reverberant room with ventilation noise and saved in a compressed format. For this illustration we used recordings from a research database. Procedures are described for simulating the recording conditions of the suspect and offender samples to convert recordings taken from the database. A pair of offender and suspect condition recordings of one speaker was selected as mock offender and suspect samples, respectively, to stand in place of the speakers on the actual casework recording.

We compared the validity and reliability of a forensic-voice-comparison system incorporating feature warping (Pelecanos and Sridharan, 2001) using Gaussian cumulative distribution function matching, probabilistic feature mapping (PFM; Mak et al., 2007), and feature-domain nuisance attribute projection (NAP; Campbell et al., 2008), as well as combinations thereof. While substantial improvements in validity were observed for all techniques, reliability deteriorated. The best performance was obtained by a combination of feature warping and probabilistic feature mapping.

The presentation will include an illustration of how we incorporated the combined feature warping and probabilistic feature mapping compensation method into our forensic-voice-comparison system, the results from testing validity and reliability of this system, and a demonstration of the evaluation of the likelihood ratio for the mock offender and suspect samples.

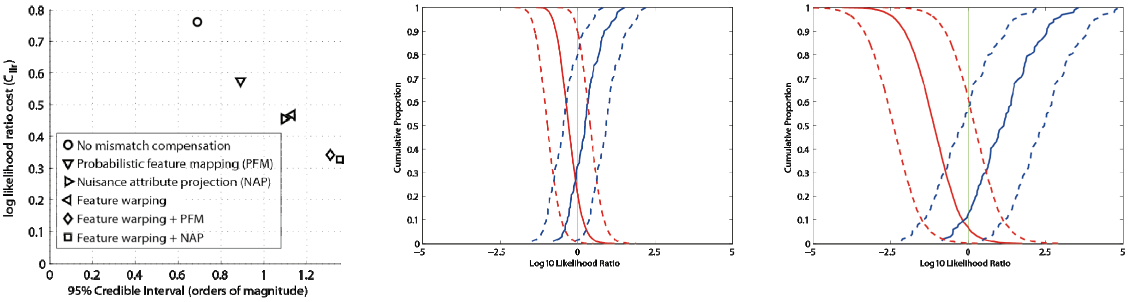


Figure 1 Measures of validity (C_{llr}) and reliability (*log*10 95% credible interval) for systems without and after incorporating mismatch compensation techniques (left); Tippett plots of the system without mismatch compensation (middle) and the system incorporating feature warping and probabilistic feature mapping (right). Solid lines represent likelihood ratios obtained from tests of the system, and dashed lines represent the 95% credible interval.

References

Campbell, W., Sturim, D., Torres-Carrasquillo, P., and Reynolds, D. (2008). A comparison of subspace feature-domain methods for language recognition. Proc. Interspeech, pp. 309–312.

Mak, M.-W., Yiu, K.-K., and Kung, S.-Y. (2007). Probabilistic feature-based transformation for speaker verification over telephone networks. Neurocomputing 71, 1-3, 137–146.

Pelecanos, J. and Sridharan, S. (2001). Feature warping for robust speaker verification. Proceedings of Odyssey 2001: The Speaker Recognition Workshop, pp. 213–218.

Eligible for student award

Speaker discrimination based on 'facewear speech'

Natalie Fecher and Dominic Watt
Department of Language and Linguistic Science, University of York, York, United Kingdom
{natalie.fecher|dominic.watt}@york.ac.uk

Introduction. Previous behavioural and neurological research has shown that speech content and speaker-specific properties of speech are processed in a mutually dependent way. It has been reported that the extraction of indexical information encoded in the speech signal – that which helps listeners to tell one speaker apart from another – depends to a significant extent upon the segmental content of an utterance (Mullennix & Pisoni, 1990; Andics *et al.*, 2007; Cutler *et al.*, 2011). Building on these findings, the present study investigates lay listeners' ability to distinguish between two unfamiliar speakers when all they have available for comparison are /Cɑː/ syllables. In this context, we examine whether some consonants possess greater speaker-discriminating potential than others. Moreover, we explore whether speaker discrimination is further complicated when the listeners' decisions are based on 'facewear speech', namely speech that has been produced while the speaker's face is disguised by a forensically-relevant face covering. The goal of this work is to extend previous research on the influence of the segmental content of an utterance on speaker discrimination, and to offer new insights into the likely effects of facial disguise on speaker discriminability.

Method. The task of 24 participants (13F, 11M, mean age 25.2) was to make timed decisions about which pair of speech samples – out of two pairs presented in each of 432 experimental trials – were produced by the same speaker ('two-interval forced-choice' procedure). The speech material was extracted from the 'Audio-Visual Face Cover' corpus (Fecher, 2012) and was highly controlled (e.g. for amplitude, interstimulus intervals, and the occurrence of a response bias). It consisted of /Cɑː/ syllables with a systematically varying onset (/p t f s m n/). These syllables were produced by four male speakers a) in a control (no facewear) condition, b) while wearing a motorcycle helmet, and c) with a piece of tape adhered across their mouths.

Results and discussion. In total, 78.2% (*SD* = 5.5) of all speaker discriminations were correct. The listeners were able to distinguish between the speakers significantly better than chance level (50%), even under the degraded listening conditions caused by the helmet and tape (*p*s <.001). Repeated-measures ANOVA revealed a significant main effect of facewear [$F(2,46) = 234.27$, $p < .001$, $\eta_p^2 = .91$] and consonant [$F(5,115) = 9.54$, $p < .001$, $\eta_p^2 = .29$] on response accuracy, as well as a significant main effect of facewear on response time [$F(1,31) = 32.75$, $p < .001$, $\eta_p^2 = .59$]. In comparison to the near-ceiling performance achieved by the listeners in the control condition (92.6%), response accuracy dropped by 18% in the helmet and 25% in the tape condition. The reduced proportion of correct responses in the two facewear conditions, along with the significant delay in response (*p*s < .001), indicate that speaker discrimination became more difficult for the perceiver – and correspondingly more error-prone – when facewear was involved in the task. Furthermore, the consonantal content of the test syllables was found to impact quite considerably on speaker discriminability. This implies that some consonants provided more speaker-specific cues that led to successful speaker discrimination than others. Further statistical evaluation and detailed auditory/acoustic analysis of the test material provided evidence that facewear modified the articulatory and acoustic properties of speech both on the segmental and suprasegmental levels. In addition, some of the facewear-induced changes to the perceptual properties of speech (see also Fecher & Watt, 2013) appeared to manifest themselves in a speaker-specific manner (i.e., some speakers seem to have been more resistant to 'facewear effects' than others).

References

Andics, A., McQueen, J. M. & van Turennout, M. (2007). Phonetic content influences voice discriminability. *Proceedings of the 16th International Congress of Phonetic Sciences* (ICPhS), Saarbrücken, Germany, August 6–10, 2007, pp. 1829–32.

Cutler, A., Andics, A. & Fang, Z. (2011). Inter-dependent categorization of voices and segments. *Proceedings of the 17th International Congress of Phonetic Sciences* (ICPhS), Hong Kong, China, August 17–21, 2011, pp. 552–555.

Fecher, N. & Watt, D. (2013). Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions. *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing* (AVSP), Annecy, France, August 29–September 1, 2013, pp. 81–86.

Fecher, N. (2012). The ‘Audio-Visual Face Cover Corpus’: Investigations into audio-visual speech and speaker recognition when the speaker’s face is occluded by facewear. *Proceedings of the 13th Annual Conference of the International Speech Communication Association* (Interspeech), Portland, Oregon, USA, September 9–13, 2012.

Mullennix, J. W. & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics* **47**(4), pp. 379–390.

Eligible for student award

Perceptual voice similarity of related speakers: telephone and microphone recordings

Hanna S. Feiser, Christoph Draxler
Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität, Munich, Germany
{feiser|draxler}@phonetik.uni-muenchen.de

Introduction

In forensic casework it is necessary to deal with similar male voices on a daily basis. Thereby some voices are more similar than others (e.g. Jessen 2012, Rose 2002). A perception experiment carried out by Feiser (2012) showed that naïve listeners are able to distinguish between related and non-related speakers. The aim of the present perception study is to test whether naïve listeners are able to identify voices of speaker pairs from telephone and microphone recordings and whether voices from related speakers are confused more often.

Methods

Recordings were obtained from ten pairs of brothers between the age of 19 and 31 and all were speakers of central Austro-Bavarian. The speakers read the *German 100 Berlin sentences* in a sound-attenuated booth. Recordings ran over two Nokia mobile phones and two *Neumann TLM 103* microphones at the same time. The perception experiment was conducted using *Percy web-experiments* and took about 20 minutes per subject. A voice identification task with no repetitions was presented to 122 listeners (64 female, 58 male) between the age of 19 and 64. 30 listeners participated in the phonetic lab and the remaining 92 attended the experiment online from different locations. Ten speaker pairs were presented to the listeners in ten separate blocks. Firstly, every block contained a training period where four stimuli (two from each speaker) were represented and the listener had to memorize a coloured symbol for each of the two speakers. Secondly, in the test period 16 stimuli were represented – one at a time from each speaker. Afterwards, listeners had to decide which speaker spoke each stimulus by clicking on the matching coloured symbol.

Results

Correct identification of the twenty speakers by their voices was significantly above chance. Listeners correctly identified the speakers in 88% of all instances (for details see Figure 1). A general linear mixed model with correct identification as the dependent variable, relation and recording type as independent variables plus listener and speaker as random factors showed that stimuli over microphone were identified significantly better than stimuli over telephone. Additionally, the number of false identifications was greater for brother pairs than for pairs with two unrelated speakers.

Discussion and conclusion

The findings clearly show that naïve listeners are able to identify speakers after a short familiarization. Results indicate that siblings' voices and voices over telephone are more often confused. Therefore, it seems that the voices of family members are perceptually more similar than those of unrelated speakers (e.g. Nolan 2009). In the present study listeners had the opportunity to use all acoustic information available in the speech signal. Results suggest that when listeners could not rely on dialect features (speakers came from the same dialect area), they had to use different features. Previous acoustic analysis of mean F0, vowel formants and articulation rate of the same 20 speakers revealed that those features seemed not to be responsible for the perceptual similarity. This raises the question of what is responsible for the similarity.

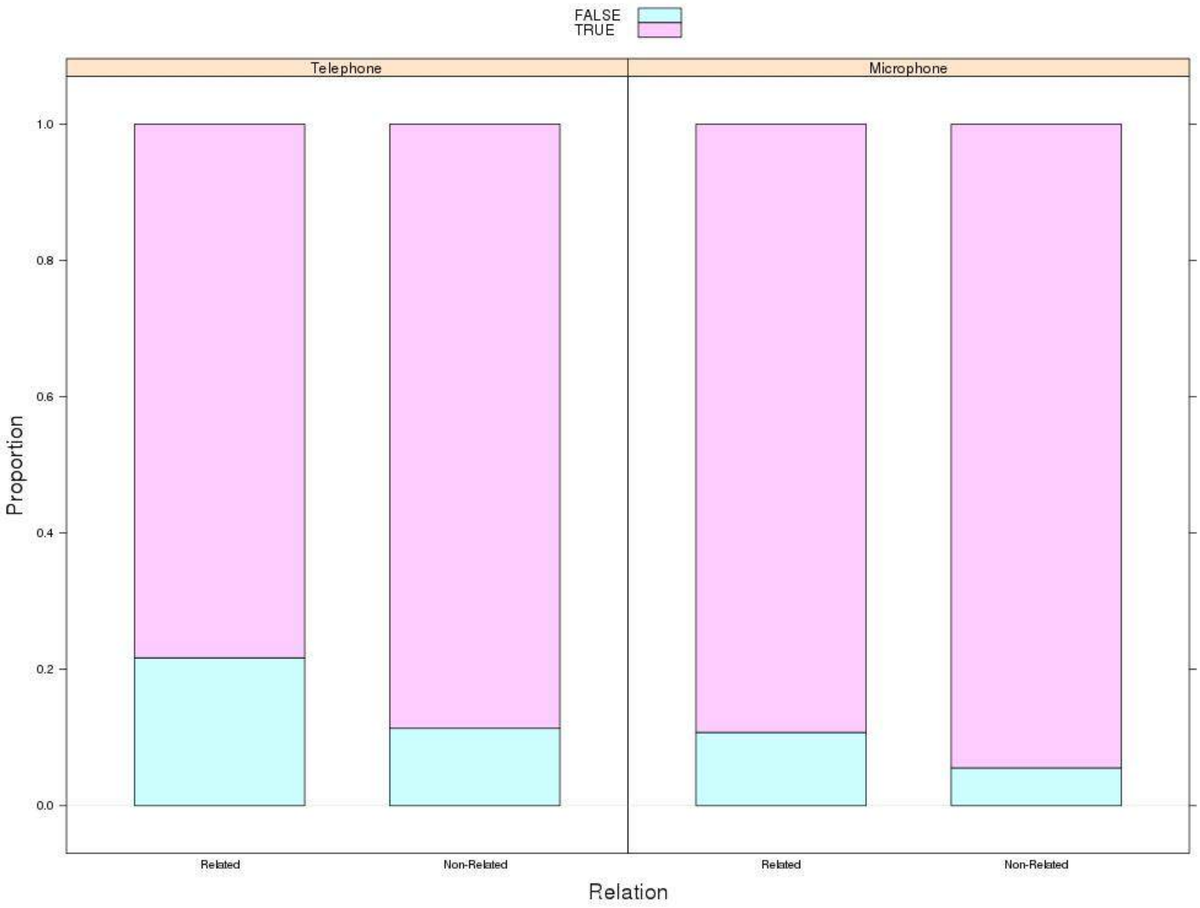


Figure 1 Proportion of true (pink) and false (turquoise) identifications separately for recording type (left box: telephone, right box: microphone) and separately for relation in each box (left: related speaker pairs, right: non-related speaker pairs).

References

Feiser, H.S. & Kleber, F. (2012). Voice similarity among brothers: evidence from a perception experiment. In Proc. 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) 2012, Santander, ES.

Jessen, M. (2012). Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs. Munich: LINCOM Studies in Phonetics.

Nolan, F. et al. (2009). *Voice similarity and the effect of the telephone: a study of the implications for earwitness evidence*. Full Research Report, ESRC End of Award Report, RES-000-22-2582. Swindon: ESRC.

Percy Web-experiments: http://webapp.phonetik.uni-muenchen.de/WebExperiment/

Rose, P. (2002). *Forensic speaker identification*. London: Taylor & Francis.

Content Comparison and Analysis (COCOA) of Contemporaneously Recorded Audio Material

Oscar Forth and Anil Alexander
Oxford Wave Research Ltd, Oxford, United Kingdom
`{oscar|anil@oxfordwaveresearch.com}`

The proliferation of handheld audio and video recorders and cheap data storage media in recent years has resulted in a large amount of audio and video evidence that is collected in both forensic and investigative tasks. It is also not unusual for an investigation to have several independent sources of audio, pictures or video, as in the case of the Boston marathon bombing in April 2013, where there was an appeal to the public for their recordings of the events leading up to the blasts. Searching, comparing, and extracting the relevant parts of the recordings as evidence is a time-consuming task, and can be helped significantly by automatic analysis techniques.

In Alexander et al (2012), we have presented a method for cancelling out music or television interference from forensic audio recordings using the so called, ‘audio fingerprinting’ method. The ability to ‘fingerprint’ a section of audio, based on the acoustic content present in it and to accurately time-align and ‘subtract’ the source material, allowed for significant improvement in the intelligibility of the target speech present in the audio. In this work, we extend this approach to all types of audio recording containing non-music speech or other sounds. We propose a novel method of comparing audio files using the acoustic content recorded in the files. If two or more different recordings contain the same acoustic events, it is possible to search for and identify the audio that is overlapping. This method will allow us to compare one audio file with many audio files in a directory, and provides a likelihood of match of a part or the whole of the files. A match provides the exact time of alignment between two recordings of the same event.

The proposed ‘COCOA’ method uses time correlations of the energy variation in the frequency spectrum to identify match points. The following are the three main applications of this method to forensic and investigative analyses:

- Content-based audio search: In certain forensic tasks, although a large quantity of audio or video data is analyzed, only a small section of audio is presented in evidence. However, it is sometimes necessary to provide the source recordings of various clips provided in court. This approach allows the forensic expert to search through audio recordings possibly from multiple cameras or recorders covering the same event.
- Intelligibility enhancement: Using time synchronization of a set of independent recorders, it is then possible either to use reference cancellation to reduce the effect of interfering noises or to mix devices for a better output. In Figure 1 and 2, we illustrate the exact time alignment of three recordings made using mobile telephones in a pub.
- Audio data de-duplication: During audio enhancement or speaker recognition work, many slightly modified copies of the audio or sections of the audio can be created on the expert’s workstation. By analyzing the content of the audio files, it is possible to identify and group files that contain overlapping or similar audio content.

This approach successfully extends the scope of audio content comparison beyond recordings with distinct frequency patterns like music and television, to more general recordings. The initial results obtained using the approach show good performance even when comparing relatively clean recordings with severely degraded ones (e.g. from a damaged recorder). This audio content-based comparison approach can be applied to a variety of forensic audio and video related problems.

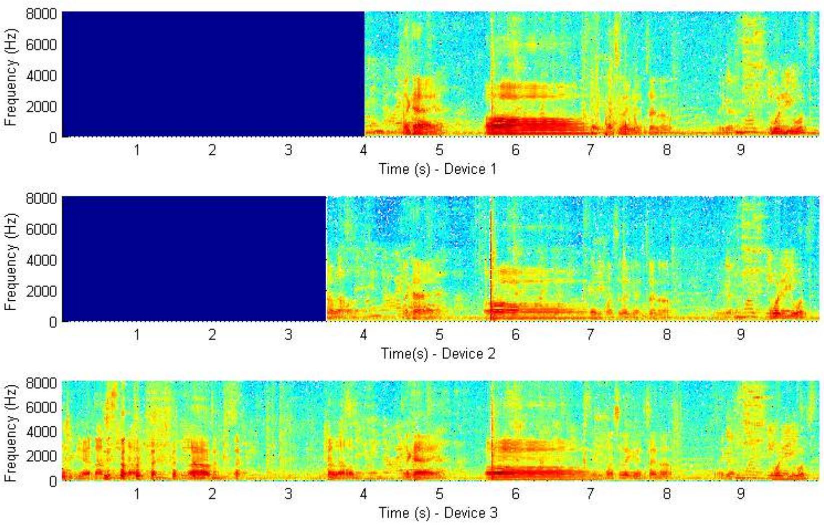


Figure 1 Time-aligned spectrograms of recordings from three independent mobile phones made in a pub environment using the COCOA method. All recorders were started at different times with device 3 started before both 1 and 2.

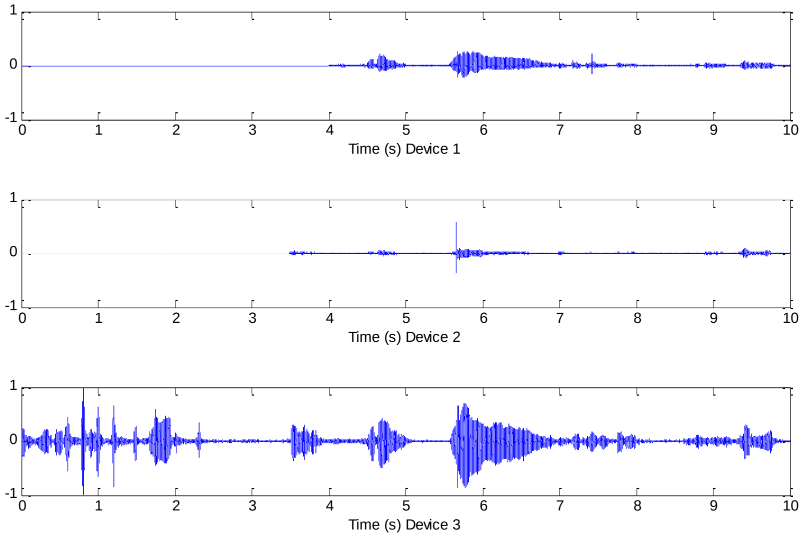


Figure 2 Time-aligned waveforms of the recordings from three independent mobile phones made in a pub environment using the COCOA method (same recordings as in Figure 1).

References

Alexander, A., Forth, O., and Tunstall, D., "Music and noise fingerprinting and reference cancellation applied to forensic audio enhancement," Audio Eng. Soc. 46th Int. Conf.: Audio Forensics, Denver, CO, pp. 29-35, June 2012

Issues in the presentation of indistinct covert recordings as evidence in criminal trials

Helen Fraser[1]
[1]*Independent Researcher, Australia*
helen@helenfraser.com.au

Covert recordings can potentially provide highly probative evidence in criminal trials. Unfortunately, due to the manner in which they are obtained, their quality is often very poor – to the extent that few words can be clearly identified by listeners with no prior knowledge of their content.

For this reason, the law in Australian and other jurisdictions allows police, in the role of so-called ‘ad hoc expert’, to transcribe indistinct covert recordings in their cases. However, since police have no real expertise in transcription (a far more skilled task than is often recognized), their transcripts are frequently inaccurate, incomplete or misleading (French & Harrison 2006).

The law seeks to mitigate this problem by requiring the jury to be cautioned that they should use the transcript only as an aid, relying on their own ears to decide what is actually said in the recording.

This paper briefly summarizes results of two sets of experiments (Fraser & Stevenson 2014; Fraser et al. 2011) which indicate this caution is unrealistic, by showing it is quite possible for juries to genuinely believe themselves to be relying on their own ears, while yet being demonstrably influenced (primed) by an inaccurate transcript.

It goes on to discuss several recent cases, suggesting it is not only juries that can be primed in this way, and showing how the current system has the potential to allow substantial miscarriages of justice.

Finally, the paper outlines efforts that have recently been made, with limited success, to bring about reform in the Australian legal system’s handling of indistinct covert recordings, and discusses some possible ways forward in the quest to ensure that, before being admitted as an aid to perception, transcripts of forensic audio are verified by appropriately qualified experts, with reference not just to what can be heard, but to acoustic phonetic evidence.

References

Fraser, H. & Stevenson, B., 2014. The power and persistence of contextual priming: more risks in using police transcripts to aid jurors' perception of poor quality covert recordings. *International Journal of Evidence and Proof*, (18), pp.205–229.

Fraser, H., Stevenson, B. & Marks, T., 2011. Interpretation of a Crisis Call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech Language and the Law*, 18(2).

French, P. & Harrison, P., 2006. Investigative and evidential applications of forensic speech science A. Heaton-Armstrong et al., eds. *Witness Testimony: Psychological, Investigative and Evidential Perspectives. Oxford: OUP*, pp.247–262.

The correlation structure of speech parameters in Southern Standard British English

Erica Gold[1] *and Vincent Hughes*[1]

[1]*Department of Language and Linguistic Science, University of York.*
{erica.gold|vh503}@york.ac.uk

The research presented in this paper is funded by a research grant from the International Association of Forensic Phonetics and Acoustics. This work builds upon a previous pilot study (Gold and Hughes 2012) and explores the correlation structure of spontaneous speech from a sociolinguistically-homogeneous set of speakers using a series of segmental, suprasegmental and linguistic parameters.

Data were extracted from a subset of speakers from the Dynamic Variability in Speech (DyViS) database (Nolan et al., 2009) and consist of:

- midpoint F1, F2 & F3 values for FLEECE (/iː/), TRAP (/a/), & NORTH (/ɔː/)
- midpoint F1, F2 & F3 values hesitation markers UM and UH
- dynamic F1, F2 & F3 values for PRICE (/aɪ/)
- long-term formant distributions (LTFD) F1-F4
- mean and standard deviation of fundamental frequency (f0)
- mean articulation rate (AR)
- voice onset time (VOT) for word-initial /t/ and /k/
- click rate (the number of velaric ingressive stops per minute)

Mean values were calculated for each speaker for each element of each variable, and a correlation matrix generated based on pairwise Spearman correlation coefficients. Pairwise correlation tests were conducted for each individual speaker and patterns compared with those of the group. Finally, Euclidean distances between variables were generated based on speaker means for each element using multidimensional scaling, as a means of developing a graphical model for all of the linguistic-phonetic variables analysed.

In terms of group patterns, a number of theoretically predictable correlations were found. There is a high degree of dependence between mean F3 values across all of the tested vocalic parameters. Similarly, a negative correlation was found between mean VOT of /t/ and mean AR. Both of these correlations are predicted by linguistic theory. The between-speaker correlation tests also revealed non-significant relationships between parameters which were expected to be correlated (f0 and F1), as well as unexpected significant correlations, such as that between mean click rate and LTFD2 (p=0.028). Interestingly, when considering patterns of correlation across elements of the same phoneme such as that between F2 and F3 of UM, a strong positive correlation was found for group means, despite some sets of within-speaker values displaying no correlation or even a negative correlation between F2 and F3.

The results highlight the overall complexity of the correlation structure of linguistic-phonetic variables as well as the extent to which this complexity is predicted by phonetic theory and the degree of agreement across within- and between-speaker correlations. The implications for combining analyses of individual speech variables into an overall assessment of the strength of evidence will be explored for both LR-

and non LR-based forensic speaker comparison.

References

Gold, E. and Hughes, V. (2012) Defining interdependencies between speech parameters. Poster presented at *the Bayesian Biometrics for Forensics (BBfor2) Summer School in Forensic Evaluation and Validation*, Universidad Autonoma de Madrid, Madrid. 18th - 21st June 2012.

Nolan, Francis, McDougall, Kirsty, de Jong, Gea and Hudson, Toby. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, 16(1): 31–57.

An exercise in calculating numerical likelihood ratios and the practicalities of their implementation

Erica Gold[1] *and Peter French*[1,2]

[1]*Department of Language and Linguistic Science, University of York, York, United Kingdom*
erica.gold@york.ac.uk

[2]*J P French Associates, York, United Kingdom*
peter.french@jpfrench.com

In recent years, there have been calls for improvements in the quality of forensic evidence by a number of legal and government bodies. It has been argued that all areas of forensic science need to be more transparent, that forensic examinations should be based on validated methodologies, and that the results should be replicable and expressed in quantitative terms (U.S. National Research Council, 2009; House of Commons' Northern Ireland Affairs Committee, 2009; Law Commission of England & Wales, 2011).

To heed these calls for improvement, a considerable amount of research in forensic speaker comparison has been devoted to the application of the numerical likelihood ratio framework (Morrison 2009). The research presented in this paper serves as an exercise in calculating numerical likelihood ratios for a linguistically-homogeneous population of 100 male, Southern Standard British English speakers (Nolan et al. 2009). This paper considers the discriminant power of four parameters (described as good speaker discriminants by experts in Gold and French 2011) in combination, while evaluating the practicalities of the numerical likelihood ratio framework for forensic speaker comparison casework.

The four parameters analyzed are articulation rate, fundamental frequency, long-term formant distributions, and the incidence of clicks (velaric ingressive plosive). Three of the parameters (clicks are excluded owing to the special difficulties they pose for statistical modeling) are combined into an overall likelihood ratio, where the combined calibrated system achieves an EER of .0554 and a Cllr of 0.2831. These results are equivalent to those achieved using a highly developed ASR on the same data, and could undoubtedly be improved upon further by the incorporation of more parameters into the overall package.

The exercise of calculating numerical likelihood ratios revealed a number of difficulties that surround the framework and its application to forensic speaker comparisons. Five prominent difficulties will be discussed in turn: subjective elements of the methodological process, delimiting the relevant reference population, availability of population statistics, lack of models available to calculate LRs, and appropriate procedures for the combination of parameters. The findings of this research are intended to promote discussion on the practical use of numerical likelihood ratios in forensic speaker comparison casework.

References

Gold, E. & French, P. (2011) International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* 18(2): 293-307.

House of Commons Northern Ireland Affairs Committee (2009). *Cross-border Co-operation between the Governments of the United Kingdom and the Republic of Ireland: Second Report of Session 2008–2009*. London: The Stationery Office.

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. *Science and Justice*, 49: 298-308.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington D.C.: The National Academic Press.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1): 31-57.

The Law Commission (2011). *Expert Evidence in Criminal Proceedings in England and Wales*. London: The Stationery Office.

Dysphonic Voice Detection for Speakers' Biometry

Pedro Gómez[1]*, Luis M. Mazaira*[1]*, Agustín Álvarez*[1]*, and Eugenia San Segundo*[2]

[1]*Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain*
pedro@fi.upm.es
[2] *Universidad Internacional Menéndez Pelayo (UIMP), Madrid, Spain*
eugeniasansegundo@gmail.com

Phonation distortion leaves relevant marks in a speaker's biometric profile. Dysphonic voice production may be used in the biometrical speaker characterization. In the present paper phonation features derived from the glottal source (GS) parameterization after the vocal tract inversion is proposed for dysphonic voice characterization in Speaker Verification tasks (Gómez, 2012). Phonated speech segments from a telephonic database of 100 male speakers (Khoury, 2013) are combined in a 10-fold cross-validation task to produce the set of quality measurements exposed in the templates of Fig. 1. Shimmer, mucosal wave correlate, vocal fold cover biomechanical parameter unbalance and a subset of the GS cepstral profile produce accuracy rates as high as 99.57 for a wide threshold interval (62,08-75.04%). An Equal Error Rate of 0.64 % can be granted. Possible applications are Speaker Verification and Dysphonic Voice Grading.

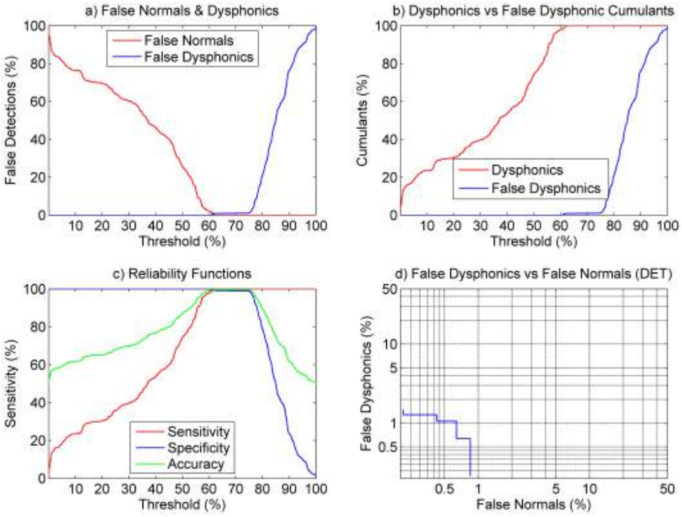


Figure 1 a) False detection rate in terms of detection threshold. b) Dysphonic and False Dysphonic Detection Cumulants (Tippett Plots). c) Merit figures: Sensitivity, Specificity and Accuracy. d) Detection Error Trade-off curves (Martin, 1997).

References

Gómez, P., et al. (2012). Distance Metric in Forensic Voice Evidence Evaluation using Dysphonia-relevant Features. *Proc. of the VI Meeting of Biometric Recognition of Persons*, Ed. Universidad de Las Palmas de Gran Canaria, pp. 169-178.

Martin, A., et al. (1997). The DET Curve in Assessment of Detection Task Performance. The National Institute of Standards and Technology, Gaithersburg, MD.

Khoury, E., Mazaira, L.M., et al. (2013). The 2013 Speaker Recognition Evaluation in Mobile environments. Proc. of the 6th IAPR International Conference on Biometrics, Madrid, Spain.

Exploring long-term formants in bilingual speakers

Willlemijn Heeren[1], *David van der Vloed*[2], *and Jos Vermeulen*[2]
[1]*Leiden University Centre for Linguistics, Leiden University, The Netherlands*
w.f.l.heeren@hum.leidenuniv.nl
[2]*Netherlands Forensic Institute, The Netherlands*
{j.vermeulen,d.van.der.vloed}@nfi.minvenj.nl

Introduction

Long-term formants (LTFs) have been forwarded as a useful feature in forensic speaker comparison (e.g. Nolan and Grigoras, 2005; Gold e.a., 2013). LTFs are assumed to be independent of individual speech sounds (Nolan and Grigoras, 2005), and earlier data support the conclusion that LTFs are language-independent (Jessen, 2010). The latter author called for more research to validate this claim, which is underlined by the finding that different speaking styles do affect LTFs (Moos, 2010). We explored if the language a bilingual is speaking influences LTFs, using forensic intercepted telephone speech (Van der Vloed et al., 2014).

Method

Recordings from twelve, male bilingual speakers of Dutch and Turkish were selected from the NFI-FRITS database (Van der Vloed et al., 2014). Recordings were pre-processed using Praat (Boersma and Weenink, 2013) to create 10-second wave files (per language and per speaker) that only included vocalic parts, following the procedure in Moos (2010). Wave files were longer than the six seconds proposed by Moos (2010) given the nature of our database: background noise and low quality may interfere with formant estimations. The first through third formant values were extracted using WaveSurfer 1.8.8p4. Only LTF2 and LTF3 were analyzed, as LTF1 is too close to the lower cutoff frequency of the phone bandwidth (see Byrne and Foulkes, 2004).

Results and discussion

To investigate if LTFs differed within speakers, but between languages, paired samples t-tests were run on both the LTF2 and LTF3 means and standard deviations. Mean LTFs within speaker and between languages did not significantly differ. Across speakers, mean LTF2 was 1418.3 Hz for Dutch and 1417.6 Hz for Turkish, and mean LTF3 was 2449.9 Hz for Dutch and 2453.8 Hz for Turkish. Standard deviations showed a difference for LTF2 ($t(11) = 2.35$, $p = .039$), but not for LTF3.

As a first comparison of the LTF distributions (LTFDs), formant histograms were compared using the Kolmogorov-Smirnov (KS) distance, either between languages within a speaker, or between speakers within a language. The KS distance is the largest absolute difference between two cumulative sample distributions. This descriptive analysis gave smaller distances for the within-speaker, between-language comparisons ($N = 12$) than for the between-speaker, within-language comparisons ($N = 66$ per language). According to Mann-Whitney-U tests, within-speaker distances for LTFD2 were smaller than between-speaker distances ($Z = -3.4$, $p = .001$), and a trend in the same direction was found for LTFD3 ($Z = -1.9$, $p = .063$).

Results are in line with previous claims that LTFs are comparable between languages, when spoken by the same speaker, and differences seem to be larger when comparing between speakers. This ties in with studies showing that LTFs may be useful in forensic speaker comparison. We aim to discuss our experiences of working with forensic speech materials, and investigate if the 10 s samples were sufficiently long for LTF estimation on such data.

References

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.42, retrieved 2 March 2013 from http://www.praat.org/.

Byrne, C. and Foulkes, P. (2004). The ‘mobile phone effect’ on vowel formants. *Journal of Speech, Language and the Law* **11***,* 83-102.

Gold, E., French, P. and Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics*, Vol. 19.

Jessen, M. (2010). *Workshop Langzeitformantenanalyse*. BKA, Wiesbaden, 28 April 2010.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician* **101***,* 7–24.

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

Van der Vloed, D. L., Bouten, J. S. and Van Leeuwen, D.A. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments *Proceedings of Odyssey Speaker and Language Recognition Workshop 2014*, Joensuu, Finland, June 16-19, 2014, pp. 6-13.

Eligible for student award

Inter-speakers variability of intensity levels across syllables

Lei He and Volker Dellwo
Phonetics Laboratory, University of Zurich, Zurich, Switzerland
{lei.he|volker.dellwo}@uzh.ch

People can easily be identified by their voice. Apart from information about the speaker's linguistic and socioeconomic background, the anatomical individualities of the speech organs and vocal tract, as well as the idiosyncratic control over the muscular movements of speech organs are fundamental to speaker idiosyncrasy in the speech signal (Dellwo et al., 2007). Such individual differences could result in speaker-specific pulmonic and sub-glottal pressure fluctuations. As a result, energy distributions in the acoustic signal could be idiosyncratic as well. We hypothesize that this speaker-specific energy distribution in the speech signal could be captured by measuring the intensity level variability across syllables in the utterances.

Based on the influential speech rhythm metrics (Ramus et al., 1999; Grabe & Low, 2002; Dellwo, 2006), which also yielded fair success in forensic voice comparison research (Dellwo et al., 2012; Leemann et al., 2014), we developed two sets of intensity variability metrics:

- The global measures:
  - stdevM: the standard deviation of average syllable intensity levels;
  - stdevP: the standard deviation of syllable peak intensity levels;
  - varcoM: the variation coefficient of average syllable intensity levels (normalized stdevM);
  - varcoP: the variation coefficient of syllable peak intensity levels (normalized stdevP).

- The local measures:
  - rPVIm: the raw pairwise variability of adjacent mean syllable intensity levels;
  - rPVIp: the raw pairwise variability of adjacent syllable peak intensity levels;
  - nPVIm: the normalized pairwise variability of adjacent mean syllable intensity levels;
  - nPVIp: the normalized pairwise variability of adjacent syllable peak intensity levels.

We applied these metrics to the TEVOID corpus built by Dellwo et al. (2012), which currently contains 16 gender-balanced Zurich German speakers, each producing 256 read sentences and 16 spontaneously uttered sentences. An initial visualization of the raw metrics scores (please see the box plots in Figure 1) suggests that a significant factor of the speakers is very likely to be found after transforming the data into more normally distributed ones. For further research, we would like to test these metrics on degraded speech as well, and work on possible optimizations of the metrics to make them more useful in forensic applications.

References

Dellwo, V. (2006). Rhythm and speech rate: a variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), *Language and Language Processing*, 231-241, Frankfurt: Peter Lang.

Dellwo, V., Huckvale, M. and Ashby, M. (2007). How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In C. Müller (Ed.), *Speaker Identification I*, 1-20, Berlin: Springer Verlag.

Dellwo, V., Leemann, A. and Kolly, M.-J. (2012). Speaker idiosyncratic rhythm features in the speech signal. In *Interspeech*, Portland, USA.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and rhythm class hypothesis. In N. Warner & C. Gussenhoven (eds.), *Papers in Laboratory Phonology 7*, 515-543, Berlin and New York: Mouton de Gruyter.

Leemann, A., Kolly, M.-J., and Dellwo, V. (2014). Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Sci. Int.*, **238**, 59-67.

Ramus, F., Nespor, M. and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **73**, 265-292.

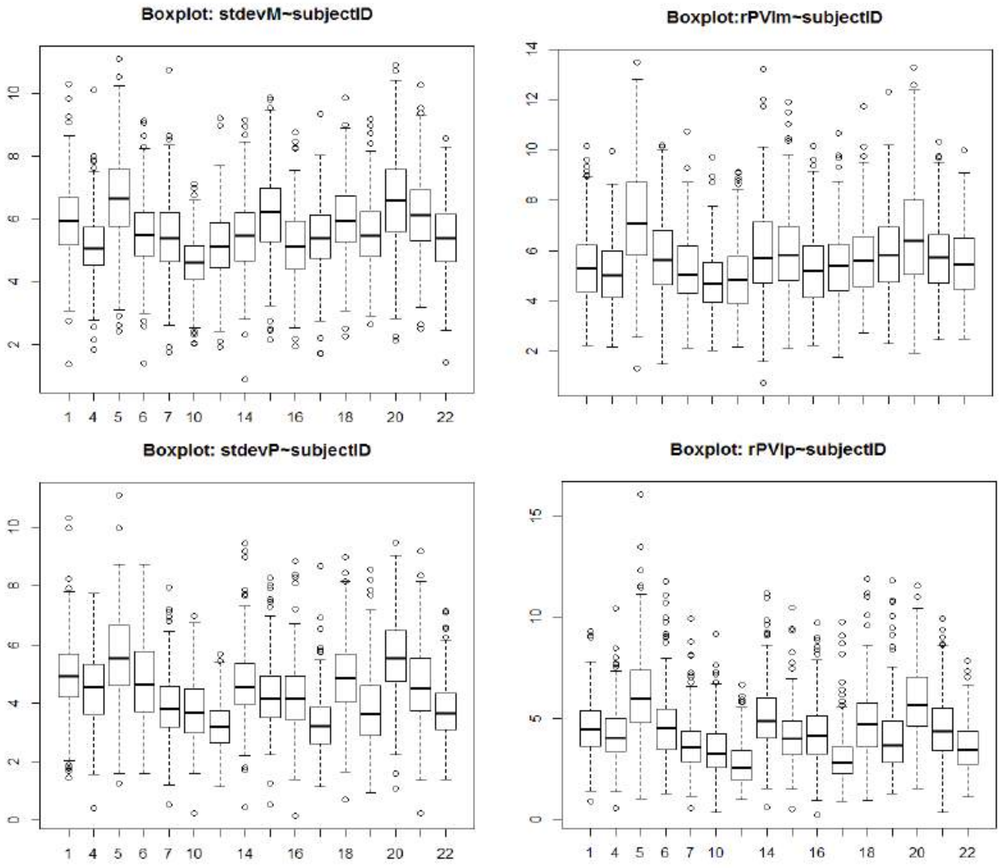


Figure 1 Box plots of the stdevM, rPVIm, stdevP and rPVIp scores of the read sentences in the TEVOID corpus.

‘Easy straight upsay hospey’ – the forensic decryption of Pig Latin

Allen Hirson[1], Lucy Dipper[1], Johan Verhoeven[1]

[1]*Language & Communication Science, City University London, UK*
A.Hirson@city.ac.uk

Abstract

The analysis of evidential speech recordings may be made more complex where it includes argots or the interweaving of different languages, intentionally or otherwise concealing what is being communicated. More obviously employed as a code or cryptolect, are language games or ‘ludlings’. Often originating as children’s secret languages, these ludlings are found across the world, sometimes doubling up as an encryption device in the criminal underworld whilst simultaneously serving to reinforce a shared group identity. Whereas many ludlings have been described in the literature, Pig Latin has received relatively little attention and is often mistakenly conflated with a much older ‘back slang’. Despite being known across a broad swath of the English-speaking world, it is demonstrated through analysis of an prison telephone recording that Pig Latin can nevertheless work surprisingly effectively as a code. The effectiveness of encryption and the reciprocal difficulty of decryption derived in part from the embedding the Pig Latin in the substrate language. This created problems in identifying the boundaries between languages, locating Pig Latin word boundaries, and patterns of lenition of Pig Latin in connected speech. Other observed patterns in the Pig Latin encryption was that it was largely (84%) restricted to ‘content’ words, and 79% were single syllable words. There was also some weak evidence that conversion into Pig Latin may be suppressed by words lacking a syllable onset - except where Pig Latin formed concatenated phrases. The concentration within content words is consistent with Pig Latin’s role as a code, even if it also serves to (re-)affirm group identity. Taking all these factors into consideration, successful decoding was achieved by application of the Pig Latin generative rule in reverse with some adjustments made for the handling of unstressed syllables in polysyllabic words.

Selected References

Barlow, J. (2001). "Individual differences in the production of initial consonant sequences in Pig Latin." *Lingua*, **111**:667-696.

Cowan, N. (1989). "Acquisition of Pig Latin: A Case Study." *Journal of Child Language,* **16** (2): 365-386.

Devlin, A. (1996). “Prison Patter: A dictionary of Prison Words and Slang.” Waterside Press, Winchester.

Laycock, D. (1972). “Towards a Typology of Ludlings, or Play-Languages.” *Linguistic Communications: Working Papers of the Linguistic Society of Australia* **6**: 61-113.

Looser, D. M. F. (2001) “Boobslang: A lexicographical study of the argot of New Zealand prison inmates in the period 1996-2000.” PhD thesis, University of Canterbury, New Zealand. (http://ir.canterbury.ac.nz/handle/10092/4789, downloaded 30th April 2014.

Marshall, D. T. (1894). “Secret Language of Children.” *Science*, **23** (572): 39.

Nevins. A & Vaux, B. (2003). "Underdetermination in language games: Survey and analysis of Pig Latin dialects." Linguistic Society of America Annual Meeting, Atlanta.

The effects of voice disguise on f0 and on the formants

Ingrid Hove and Volker Dellwo
Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland

ingrid.hove@uzh.ch; volker.dellwo@uzh.ch

Voice disguise is a serious problem for forensic speaker identification. In order to help provide solutions to deal with disguised (or possibly disguised) voices we aim at finding out which acoustic characteristics change and which remain consistent in different disguise conditions. For the characteristics which are affected by voice disguise, the aim is to find out whether these changes are systematic, i.e. whether they always go in the same direction for a specific disguise condition or not.

Previous research in the temporal domain has shown that certain durational characteristics are idiosynchratic throughout different disguise conditions (Hove & Dellwo 2012). In the present study, we focus on effects of different kinds of voice disguise in typical frequency-domain based speaker specific characteristics like average fundamental frequency and formants.

The corpus we recorded contains read speech of 12 speakers of Zurich German. Every speaker reads 24 translated sentences from the Bamford-Kowal-Bench corpus (Bench/Kowal/Bamford 1979), 12 sequences of nonsense words of the type $CV^1CV^2CV^3$ with each word pronounced three times, plus a well-known Swiss German nursery rhyme. The disguise conditions are two types of prosodic modification, namely high-pitched voice and low-pitched voice, as well as four types of articulatory obstruction: speaking with a pencil in the mouth, speaking with a lollipop in the mouth, speaking with a pinched nose and speaking with a hand in front of the mouth. For all sentences we compare the fundamental frequency (mean pitch), the standard deviation and the minimum and maximum of the fundamental frequency.

For the prosodic modification conditions, first results show that the speakers differ in their ability to modify their pitch: when speaking in a high- or low-pitched voice certain speakers succeed well in raising or lowering their voice whereas others only show small differences to their normal speaking voice. We are also looking at how consistently the speakers can keep up this modified pitch. Furthermore, the effect of the disguise conditions on the formants will be examined. For this part of the analysis, the focus will lie on the nonsense words of the type [paːpiːpuː] or [xiːxyːxuː].

The comparison of our results to the findings of other studies on voice disguise such as Künzel (2000), Perrot et al. (2007), Moosmüller (2001), Eriksson and Wretling (1997), or Masthoff (1996) should expand our understanding of the effects of voice disguise on the speech.

References

Bench J., A. Kowal and J. Bamford. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. Brit J Audiol. 13, 108–112.

Boersma, P. and D. Weenink. Praat: Doing phonetics with computers, www.praat.org, accessed 31 Mar 2014.

Eriksson, A. and P. Wretling. (1997). “How flexible is the human voice? – A case study of mimicry”, Eurospeech 97(2), 1043-1046.

Hove, I. and V. Dellwo. (2012). *The effect of articulatory obstruction on temporal characteristics of speech*. Abstract presented at IAFPA 2012, Santander/Spain.

Künzel, H. (2000). “Effects of voice disguise on speaking fundamental frequency”, Forensic Linguistics, 7(2):150- 179, 2000

Masthoff, H. (1996). “A report on a voice disguise experiment”, Forensic Linguistics, 3(1), 160-167.

Moosmüller, S. (2001). “The influence of creaky voice on formant frequency changes”, The International Journal of Speech, Language and the Law, 8/1, 100-112.

Perrot, P., G. Aversano & G. Chollet (2007). "Voice Disguise and Automatic Detection: Review and Perspectives", Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.): WNSP 2005, LNCS 4391, Berlin: Springer, pp. 101 – 117.

Using the smartphone application ‘Voice Äpp’ to collect speech population data: implications for forensic phonetics

Ingrid Hove[1], Adrian Leemann[1], Marie-José Kolly[1], Volker Dellwo[1], Jean-Philippe Goldman[2], Ibrahim Almajai[2], Daniel Wanitsch[3]

[1] Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland
`ingrid.hove@uzh.ch, {adrian.leemann|marie-jose.kolly}@pholab.uzh.ch, volker.dellwo@uzh.ch`

[2]Department of Language Sciences, University of Geneva, Switzerland
`{jean-philippe.goldman|ibrahim.almajai}@unige.ch`

[3]iBros.ch LTD, Aarau, Switzerland
`dani@ibros.ch`

The smartphone application *Voice Äpp*, which is currently in development, aims at providing its users with scientifically sound phonetic and dialectological information on their dialect and their voice and on general aspects of speech. For forensics, the users' recordings provide a valuable database for extracting phonetic population data.

- The application’s "dialect profile" functionality is designed to determine users’ dialect based on their pronunciation of 15 words using automatic speech recognition (cf. Kolly & Leemann, accepted). Since in this functionality the algorithm for dialect recognition is based on Swiss German data, this part can only be used by German speaking Swiss. The other two functionalities work for all users who understand German.
- The aim of the application’s "voice profile" functionality is that users get to know characteristics of their own voice. After having recorded a given sentence in their dialect, users are shown histograms displaying their fundamental frequency and articulation rate in comparison to all of the previous users of the application.
- In the application’s "infotainment" functionality the user can learn about different aspects of speech in a playful way, for example by listening to different kinds of hearing impairments or by experiencing the "McGurk effect" (McGurk & MacDonald, 1976) and the "cocktail party effect" (Handel, 1989).

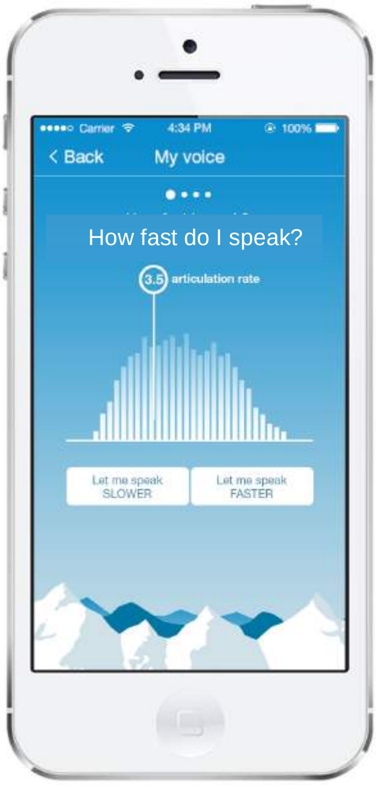


Figure 1 Screen shot showing the user's articulation rate

From a scientific point of view, *Voice Äpp* allows crowdsourcing of population data which has important implications for forensic voice comparison research. Acoustic analyses of the users' recordings will allow unprecedented insights on the areal distribution of speech signal parameters such as fundamental and formant frequencies, temporal characteristics of segments, and speaking rate. Forensic phonetic research requires population data from a large set of speakers. Until now, population statistics only exist for certain languages (English, Standard German) and typically are based on the data of around 50–100 speakers (Künzel, Masthoff & Köster, 1995; Jessen, 2007). When collecting data through crowdsourcing, certain parameters are not controllable. This disadvantage is compensated by the large amount of data we are expecting based on our experience with the predecessor application *Dialäkt Äpp* (Kolly & Leemann, accepted).

References

Boersma, P. and D. Weenink. Praat: Doing phonetics with computers, www.praat.org, accessed 31 Mar 2014.

Kolly, M.-J. & Leemann, A. (accepted). *Dialäkt Äpp*: communicating dialectology to the public – crowdsourcing dialects from the public. In: Leemann, A., Kolly, M.-J., Schmid, S. & Dellwo, V. (Eds.). *Trends in Phonetics in German-speaking Europe*, Bern/Frankfurt: Peter Lang.

Künzel, H., Masthoff, H., & Köster, J. (1995). The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. *Science and Justice*, *35*, 291–295.

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science and Justice,* 47, 50–67

McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 746-748.

Handel, S. (1989): *Listening. An Introduction to the perception of auditory events*. Cambridge, Mass.: MIT Press.

Modelling features for forensic speaker comparison

Vincent Hughes[1], *Erica Gold*[1], *Paul Foulkes*[1], *Peter French*[1,2], *Philip Harrison*[1,2], *Louisa Stevens*[1,2], *Colin Aitken*[3] *and Tereza Neocleous*[4]

[1]*Department of Language and Linguistic Science, University of York.*
{erica.gold|vh503|paul.foulkes}@york.ac.uk
[2]*J P French Associates, York.*
{ jpf|pth|lcc }@jpfrench.com
[3]*School of Mathematics, University of Edinburgh.*
c.g.g.aitken@ed.ac.uk
[4]*School of Mathematics and Statistics, University of Glasgow.*
tereza.neocleous@glasgow.ac.uk

Speech is an exceptionally complex form of forensic comparison evidence. In linguistic-phonetic forensic speaker comparison (FSC) experts typically analyse a range of segmental, suprasegmental, syntactic, lexical and linguistic features (Gold and French 2011). Such features can be continuous, discrete or even both (e.g. vowels can be analysed continuously using formant frequencies or discretely by considering the realisation of different allophones). Linguistic data can be normally and non-normally distributed, and features vary systematically within- and between-speakers according to a wide range of social, stylistic and phonological factors. Further, given that the speaker-space (Nolan 1991) is so highly multidimensional, there is considerable interrelatedness between features, some of which differ within- and between-speakers (see Gold and Hughes 2012). Such issues cause significant difficulty for the application of the numerical likelihood ratio (LR) framework to speech evidence since current formulae, which were never primarily designed to deal with linguistic-phonetic data, generally fail to account adequately for the complexity and interrelatedness of features.

To address these problems, a network has been established which brings together members of York's forensic speech science group with leading forensic statisticians. Building on Aitken and Gold (2013), the goals of the network are (i) to develop statistically appropriate models for analysing phonetic data of multiple types, and (ii) to explore the mathematical complexity of phonetic data. The collaboration will yield new statistical methodologies relevant to statisticians interested in multivariate data analysis, Bayesian modelling and Bayesian networks, as well as forensic speech scientists working on FSC research and casework.

In this paper, we will explore, in more detail, current issues with the application of the numerical LR to linguistic-phonetic FSC evidence and provide an overview of the aims of the network. We will also present preliminary results on two lines of work: (1) attempts to model and combine a subset of short vowels (KIT, DRESS, TRAP, LOT and STRUT) for 25 speakers from the DyViS corpus (Nolan et al. 2009); and (2) quantifying and modelling voice quality and vocal setting, based on multidimensional auditory vocal profile analyses (VPA; Laver 1991, 1994) of 100 DyViS speakers (Stevens, in progress).

References

Aitken, C. G. G and Gold, E. (2013) Evidence evaluation for discrete data. *Forensic Science International* 230: 147–155.

Gold, E. and Hughes, V. (2012) Defining interdependencies between speech parameters. Poster presented at *the Bayesian Biometrics for Forensics (BBfor2) Summer School in Forensic Evaluation and Validation*, Universidad Autonoma de Madrid, Madrid. 18th - 21st June 2012.

Laver, J. (1991) A perceptual protocol for the analysis of vocal profiles. In Laver, J. (ed.) *The Gift of Speech: Papers in the Analysis of Speech and Voice*. Edinburgh University Press: Edinburgh. pp. 265–280.

Laver, J. (1994) *Principles of Phonetics*. Cambridge University Press: Cambridge.

Nolan, F. (1991) Forensic Phonetics. *Journal of Linguistics* 27: 483–493.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1): 31–57.

Stevens, L. (in progress) Aspects of voice quality. PhD Thesis, University of York, UK.

Comparing MVKD and GMM-UBM applied to a corpus of formant-measured segmented vowels in German

Michael Jessen
Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany
michael.jessen@bka.bund.de

Currently, the two common methods of obtaining likelihood ratios for the purpose of system evaluations in forensic voice comparisons are the MVKD approach, which was originally proposed by Aitken & Lucy (2004), and the GMM-UBM approach, which was originally proposed within the context of automatic speaker recognition. The MVKD approach has been developed for **token based** scenarios. For example, formant frequencies are measured at the center of about ten tokens of a vowel category per recording (e.g. Morrison et al. 2011). The GMM-UBM approach has been developed for **data-stream-based** scenarios. This applies to MFCC feature vectors used in automatic speaker recognition, which are extracted as a data stream with a sampling rate of about ten milliseconds. These data-stream-based scenarios are not limited to automatic speaker recognition but can also be used with acoustic-phonetic data, for example long-term formants, where formant feature vectors are extracted in close temporal succession across vowels (Becker et al. 2008). Occasionally, one of the methods of obtaining likelihood ratios has been used across the scenarios. For example, Morrison (2011) applied both GMM-UBM and MVKD to tokenized data (diphthong contour parameters). However, using GMM-UBM on tokenized data turned out to be not always successful (Zhang et al. 2011; Rose 2013).

In the present experiment the two methods are compared in their “natural habitat”, i.e. GMM-UBM with data streams and MVKD with tokens-based data. The speech corpus used for this purpose is a mobile-phone transmitted portion of Pool 2010 (Jessen et al. 2005) in which 21 male adult speakers of the West-Central regional variety of German spoke in a spontaneous style, which was compared to them speaking in a semi-spontaneous style (Jessen et al. 2013 for further details). Recordings with net durations between 20 and 40 seconds were segmented for the vowels /I/ (short/lax i), /a/ and /@/ (schwa) and measured for F1, F2 and F3. Token-based data were extracted using the point-labeling facility of Praat (labeling a vowel at a point minimally influenced by context) and stream-based data by interval labeling (labeling a vowel from beginning to end). The label information was exported to Wavesurfer, where the formant tracking and manual correction were carried out. MVKD was applied based on the implementation by Morrison (http://geoff-morrison.net/) and GMM-UBM was applied based on VOCALISE (http://www.oxfordwaveresearch.com/j2/vocalise), including its region-conditioning tool SPARSE (Jessen et al. 2014 for examples). The likelihood scores obtained with these methods subsequently underwent calibration and fusion. Some of the results are shown in Figure 1. It shows that MVKD and GMM-UBM, when used in their “natural habitat”, have similar performance, although the results of GMM-UBM were mostly better under fusion. Figure 1 also shows that different vowels yield different patterns. For example**,** schwa has the lowest performance, probably due to its strong coarticulation, hence highest intra-speaker variation. Overall, fusing different vowels leads to improvement, but less strongly than in Morrison et al. (2011). Fusion was also applied between the data shown in Figure 1 and Long-Term Formants F1, F2, F3 (Jessen et al. 2013), which have an EER of 8.85 and C_{llr} of 0.395. However, no systematic improvement in speaker discrimination was obtained.

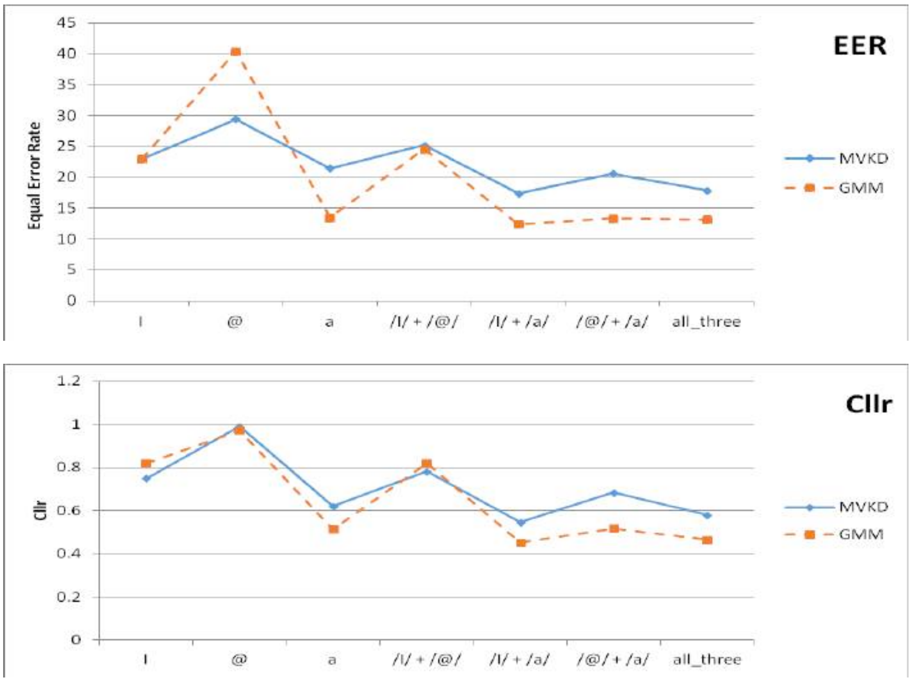


Figure 1 Equal Error rate (upper graph) and C_{llr} (lower graph) using MVKD (uninterrupted lines) and GMM-UBM (interrupted lines) for the three vowels individually (first three entries on x-axis) and fusion between different vowels (remaining entries) on vowel-segmented data from the Pool 2010 corpus.

Acknowledgement: The work of Nicola Wagner in performing vowel annotation and formant measurements and the one of Ewald Enzinger in performing MVKD analysis, calibration, fusion and related forensic-statistic procedures is gratefully acknowledged.

References

Aitken, C.G.G. and D. Lucy (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, **53**, 109–122.

Becker, T., M. Jessen and C. Grigoras (2008). Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of INTERSPEECH '08*, Brisbane, 1505–1508.

Jessen, M., O. Köster and S. Gfroerer (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, **12**, 174–213.

Jessen, M., E. Enzinger & M. Jessen (2013). Experiments on Long-Term Formant Analysis with Gaussian Mixture Modeling using VOCALISE. *Paper presented at the IAFPA Conference*, 2013, Tampa, Fl.

Jessen, M., A. Alexander and O. Forth (2014). Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proceedings of the Audio Engineering Society 54th International Conference*; London, June 12–14, pp. 28–35.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model - universal background model (GMM-UBM). *Speech Communication*, **53**, 242–256.

Morrison, G. S., C. Zhang & P. Rose (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, **208**, 59–65.

Rose, P. (2013). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language and the Law*, **20**, 77–116.

Zhang, C., G.S. Morrison and T. Thiruvaran (2011). Forensic voice comparison using Chinese /iau/. *Proceedings of the International Congress of Phonetic Sciences*, Hong Kong, pp. 2280–2283.

„www.regionalsprache.de (REDE)“ – a dialectological GIS for linguists and forensic phoneticians

Roland Kehrein [1] *and Gea de Jong-Lendle*[2]
[1]*Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg, Germany*
kehrein@uni-marburg.de
[2]*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*
gea.dejong@staff.uni-marburg.de

GIS, also known as geographical information systems, are computer systems that enable users to capture, store, analyse or query geographically referenced data. In this paper, a GIS tool is introduced that offers a range of services useful for the dialectological analysis of sound recordings involving German speakers. A demonstration is given of analysis tools that are useful for linguistic research in general and specific features that could particularly assist the forensic phonetician.

Background

The dialectological GIS tool discussed here was created as part of the regionalsprache.de project, also called the REDE-project. The goal of the “regionalsprache.de (REDE)” project in Marburg is to capture the entire variative spectrum of male speakers of different ages and social backgrounds from 150 locations across the German Federal Republic. The reason for this project was the fact that whereas the development of the different dialect varieties in German had been studied in great detail, an overview of the regiolect variety (or the regional High-German variety) or the modern regional languages as a whole, does not exist (cf. Kehrein 2006); for large parts of the Upper and West Middle German regions, developments in the dialects can be followed in precise detail over more than a century thanks to Georg Wenker. Around 1880 he conducted an extensive survey, investigating a range of aspects concerning dialectal phonology and morphology. As informants he used elementary school teachers from around 44,000 school locations from across the German Empire of the time and established the famous linguistic database that is now known as „Sprachatlas des Deutschen Reichs“.

At the time, 140 years ago, dialects dominated everyday communication in most regions of Germany. Nowadays, however, it is for a majority of speakers more common to speak a regiolect variety instead of a local dialect. Using partially the same speech materials as Wenker used at the time, the REDE study provides an overview of the current language situation in Germany concentrating in particular on inter-regional and intra-regional differences.

The speakers

At each location, three groups of speakers are examined in REDE: 1) a representative of the older generation (a so-called “NORM”), 2) two police officers as representatives of the average speaker of the modern regional language (middle generation, middle level of education and social status, communicative occupation), and 3) a representative of the potentially progressive type of speaker (17-22 years old, higher secondary education).

The speech materials

The speaking conditions recorded include read (the Northwind and the Sun) and spontaneous speech (interview/ conversation with a friend/telephone call). In addition, each speaker was asked to provide a dialect-version and their best High-German version (also referred to as their regiolect version) of the 40 sentences used in the Wenker survey.

How can the REDE database be used for forensic analysis?

Just a few examples are mentioned here:

1. All materials, that means all recordings and all the associated dialect maps, can be consulted for free. For scientific purposes materials can also be received in high quality.

2. Orthographic transcriptions exist for all recordings and phonetic transcriptions for part of the REDE data.

3. The transcripts above can be searched for words, sounds and sound combinations.

4. Geographic maps can be created online and downloaded showing the different pronunciations of particular words by region/city.

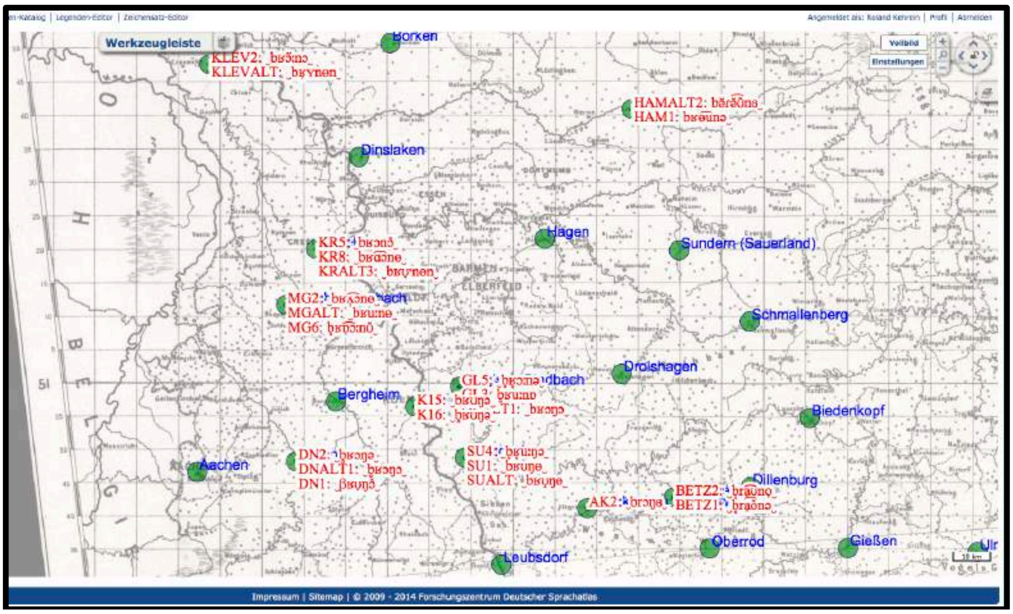


Figure 1 Example of a REDE map showing the different pronunciations of the German word `braune`(= brown). Each speaker is assigned with a code (shown in red), consisting of the abbreviation indicating the place of birth of the speaker (K=Koeln, KR=Krefeld, etc) and the generationmarker (ALT=old, JUNG=young, no marker=middle-aged)

References

Kehrein, R. (2006). Regional accent in the German language area. How dialectally do German police answer emergency calls? In: F. Hinskens (ed.). *Language Variation – European Perspectives,* 83-96. Amsterdam/Philadelphia: Benjamins.

Eligible for student award

Speaker-individual rhythmic features in both L1 and L2 speech: implications for forensic voice comparison

Marie-José Kolly, Adrian Leemann, and Volker Dellwo
Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland
{marie-jose.kolly|adrian.leemann}@pholab.uzh.ch,
volker.dellwo@uzh.ch

It is known that speakers often transfer speech rhythmical patterns from their L1 to their L2, which may affect their intelligibility (Adams, 1979; Wenk, 1985). In the present contribution we address how these L1-interference phenomena could be leveraged for forensic phonetic purposes: Do certain (speaker-individual) rhythmic characteristics remain unchanged when a speaker talks in different languages?

A number of speech rhythmic features, e.g. the percentage of voiced portions in the speech signal (Dellwo, Fourcin & Abberton, 2007), were shown to have potential for forensic voice comparison as they strongly vary between speakers but remain largely unaffected by within-speaker variability in speaking style (spontaneous vs. read) and transmission channel (hifi vs. telephone) (Leemann, Kolly & Dellwo, 2014), and by within-speaker variability when speakers disguise their voice by obstructing their articulators (Leemann, Hove, Kolly & Dellwo, submitted). The overall objective of the present contribution is to examine speech for speaker-individual rhythmic features that are independent of the language being spoken.

Our research is based on the TEVOID corpus (Dellwo, Leemann & Kolly, 2012; Leemann, Kolly & Dellwo, 2014) that contains Zurich German (L1) speech of 16 speakers and French and English (L2) speech of the same 16 speakers. Results based on 16 sentences per speaker and language showed that selected, automatically extracted rhythmic measures, e.g. the percentage of voiced portions in the speech signal, varied between speakers but remained largely unaffected by within-speaker variability in the language spoken (Kolly, Dellwo & Leemann, 2013). We have now collected more material per speaker and are currently segment-labeling this material, which will allow us to calculate a wider variety of rhythmic measures.

The present contribution reports on between- and within-speaker variability of a number of rhythmic measures, using 32 Zurich German (L1), 32 French (L2) and 32 English (L2) sentences per speaker. Based on preliminary results (cf. Kolly, Dellwo & Leemann, 2013) we expect high between- and low within-speaker variability in selected measures of speech rhythm.

In forensic voice comparison, cases occur where there is a mismatch in language between acoustic trace and comparison material. In a considerable number of forensic phonetic casework, practitioners have to make decisions about speaker identity based on speech samples where the trace material is in one language – e.g. the speaker's L1–, and the suspect material is in another language – e.g. the speaker's L2 (Herbert R. Masthoff, personal communication). This may happen, for example, when a suspect uses an L2 in order to disguise his/her voice. However, the impact of L2 speech on speaker-individual characteristics is largely unknown – this is why forensic phoneticians "should exercise particular caution if the samples for comparison are in different languages" (IAFPA Code of Practice). The present contribution is thus expected to have implications for forensic voice comparison.

References

Adams, C. (1979). *English speech rhythm and the foreign learner*. The Hague: Mouton.

Bohn, O.-S. (1998). Wahrnehmung fremdsprachlicher Laute. Wo ist das Problem? In H. Wegener (Ed.), *Eine zweite Sprache lernen. Empirische Untersuchungen zum Zweitspracherwerb*, 1–20. Tübingen: Narr.

IAFPA Code of Practice. http://www.iafpa.net/code.htm (accessed 28.04.2014).

Dellwo, V., A. Fourcin and E. Abberton. (2007). Rhythmical classification based on voice parameters. *Proceedings of the ICPhS 2007*, Saarbrücken: 1129–1132.

Dellwo, V., A. Leemann and M.-J. Kolly (2012). Speaker idiosyncratic rhythmic features in the speech singal. *Proceedings of Interspeech 2012*, Portland, OR (USA).

Leemann, A., M.-J. Kolly and V. Dellwo. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, **238**, 59–67.

Leemann, A., I. Hove, M.-J. Kolly and V. Dellwo (submitted). Testing the effect of voice disguise on suprasegmental temporal features: Corpus creation and preliminary results. Submitted to the Interspeech 2014 conference.

Lloyd James, A. (1929). *Historical introduction to French phonetics*. London: ULP.

Kolly, M.-J., V. Dellwo and A. Leemann. (2013). Speaker-idiosyncratic temporal patterns in L2 speech. Abstract presented at the annual IAFPA 2013 meeting, Tampa, FL (USA).

Taylor, D. S. (1981). Nonnative speakers and the rhythm of English. *International Review of Applied Linguistics in Language Teaching*, **19**, 221–226.

Wenk, B. (1985). Speech rhythms in second language acquisition. *Language and Speech*, **28**, 157–175.

Testing the effect of dialect imitation on suprasegmental temporal features

Adrian Leemann, Marie-José Kolly, Volker Dellwo
Department of Comparative Linguistics, University of Zurich
{adrian.leemann|marie-jose.kolly}@pholab.uzh.ch,
volker.dellwo@uzh.ch

Voice disguise is the intentional act of changing one’s voice for the purposes of falsifying identity. The German Federal Police Office (BKA) reports that more than half of forensic cases feature a form of voice disguise (Masthoff 1996). Present research distinguishes between two types of voice disguise: electronic and non-electronic voice disguise (Künzel, 2000, Masthoff, 1996, Perro et al., 2007). Previous studies have reported that voice disguise can lead to high within-speaker variability, affecting various acoustic parameters: Hollien (1977), for instance, reported high within-speaker variability in long-term spectra under various voice disguise conditions (disguise of the speaker’s own choice). Eriksson & Wretling (1997) as well as Endres et al. (1971) found that imitating another speaker is particularly achieved by adopting f0 and formant frequencies of the target speaker. High within-speaker variability makes it difficult for forensic practitioners to draw conclusions about the speakers’ identity.

The objective of this contribution is to present first results of a study that tests the effect of dialect imitation as a form of non-electronic voice disguise on suprasegmental temporal features. We focus on suprasegmental temporal features because previous research in forensic phonetics has shown that these features seem relatively robust towards different variability conditions: the amount of voiced portions in the speech signal, for instance, is affected only very little by articulatory obstructions (Leemann, Hove, Kolly, Dellwo, 2014), and the amount of vocalic portions in the signal seem to remain relatively speaker-specific even if a speaker imitates a different dialect (Dellwo et al. 2009). Leemann et al. (2014) reported that a number of suprasegmental temporal features demonstrate little within-speaker variability across different speaking styles and signal distortions.

Methodologically, we proceeded as follows: 10 speakers of Zurich German were recorded at the University of Zurich. Speakers were students and showed little to no regional accent variability in Zurich German. The data was recorded in a sound-treated booth. Each speaker read 72 sentences from the TEVOID corpus (Dellwo et al., 2012, Leemann et al., 2014). Sentences typically included 15–20 syllables and were written in Zurich German. These sentences were read by the subjects and recorded as a control. The same sentences were transliterated to Bern German and read by the same subjects for the Bern dialect imitation condition. We selected Bern German since previous research has reported differences in the suprasegmental temporal features for these two dialects (Leemann, 2012, Leemann et al., 2012). We applied the following automatic measures on the labeled corpora: (1) measures that are based on intervals between amplitude peak points of a low frequency amplitude envelope (<10Hz) and (2) measures that are derived from the amount of speech voicing in the signal. In the present contribution we will present first results and discuss these findings against the backdrop of forensic phonetics.

References

Dellwo, V., A. Leemann and M.-J. Kolly. (2012). Speaker idiosyncratic rhythmic features in the speech signal. *Proceedings of Interspeech 2012*, Portland (OR).

Dellwo, V., S. Ramyead and J. Dankovicova. (2009). The influence of voice disguise on temporal characteristics of speech. Abstract presented at the IAFPA conference 2009, 02.–05.08.2009, Cambridge, UK.

Endres, W., W. Bambach and G. Flösser. (1971). Voice spectrograms as a function of age, voice disguise and voice imitation. *Journal of the Acoustical Society of America*, **49(6)**, 1842–1848.

Eriksson, A. and P. Wretling. (1997). How flexible is the human voice? – A case study of mimicry, *Eurospeech*, **97(2)**, 1043–1046.

Hollien, H. and W. Majewski. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *Journal of the Acoustical Society of America*, **62**, 975–980.

Künzel, H. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, **7(2)**, 150–179.

Leemann, A. (2012). *Swiss German Intonation Patterns*. Amsterdam: Benjamins.

Leemann, A., V. Dellwo, M.-J. Kolly and S. Schmid. (2012). Rhythmic variability in Swiss German dialects. *Proceedings of Speech Prosody*, Shanghai.

Leemann, A., I. Hove, M.-J. Kolly and V. Dellwo (submitted). Testing the effect of voice disguise on suprasegmental temporal features: Corpus creation and preliminary results, Interspeech 2014.

Leemann, A., M.-J. Kolly and V. Dellwo. (2014). Speaker-individuality in the time domain: implications for forensic voice comparison. *Forensic Science International* **238**, 59–67.

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics*, **3(1)**, 160–167.

Perrot, P., G. Aversano and G. Chollet. (2007). Voice disguise and automatic detection: review and perspectives. In Y. Stylianou et al. (Eds.), *Progress in nonlinear speech processing 2005, LNCS 4391*,101–117.

Assessing the potential of crowdsourced ‘Dialäkt Äpp’ speech data for forensic phonetics

Adrian Leemann and Marie-José Kolly
Department of Comparative Linguistics, University of Zurich
{adrian.leemann|marie-jose.kolly}@pholab.uzh.ch

The free of charge iPhone application *Dialäkt Äpp* (Leemann & Kolly, 2013; Kolly & Leemann, in press) features the following two core functionalities: (1) users click on the pronunciation variants of 16 words and the application predicts their local dialect, (2) users record their pronunciation of the same 16 words, which are then uploaded on a server and displayed on an interactive map. The goal of the application is science communication to a broad public. The app has been downloaded by >59’000 users.

As speech scientists we are now in the position to analyze the data gathered through *Dialäkt Äpp*. With the users’ consent, we retrieve acoustic pronunciation data of 16 words for thousands of dialect speakers originating from all over German-speaking Switzerland (cf. function (2) above). Until recently, traditional methods for empirical linguistic research based their analyses mostly on small sets of speakers. The use of smartphone app technology for crowdsourcing linguistic data is relatively new: smartphone applications have hitherto been used to collect speech to train acoustic models (de Vries, Davel, Badenhorst, Basson, de Wet et al., 2014) or to document endangered languages (Iwaidja Inyman Team, 2012).

The crowdscourced speech data from *Dialäkt Äpp* allows for the collection and analysis of a number of speech signal parameters in order to create large-scale population statistics. In the field of forensic phonetics, such population statistics only exist for certain languages and typically feature <150 speakers (Künzel, Masthoff & Köster, 1995; Jessen, 2007). A preliminary analysis of *Dialäkt Äpp* recordings of 115 users from Bern (city) and 205 users from Zurich (city) revealed that Bern SwG speakers speak significantly slower than Zurich SwG speakers. For 6 disyllabic words per speaker we measured the temporal duration between the two vowel onsets. We call this vowel-onset-to-vowel-onset measure *durVonVon* (see Figure 1). In theory, this measure is motivated by Allen’s (1972) findings that vowel onsets represent perceptually prominent centers of a syllable.

Figure 1 shows the boxplots of the two dialects’ *durVonVon* values. The longer the temporal duration between the two vowel onsets, the slower the articulation rate. The values between the two dialects are significantly different: the durational information contained in a few words alone discriminates the two dialects (cf. Leemann, Kolly & Dellwo, 2014). In the present contribution, we will use this example of articulation rate differences to illustrate the potential of *Dialäkt Äpp* speech data for forensic phonetic purposes.

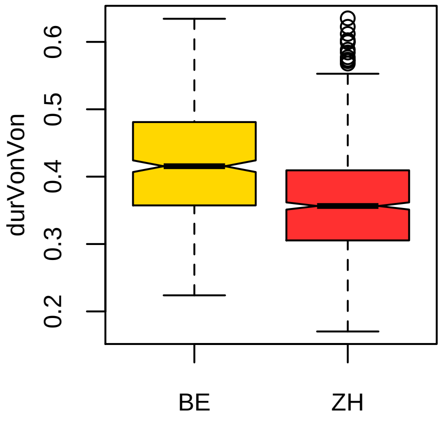


Figure 1: *Boxplots of the dialects’ durVonVon values.*

References

Allen, G. D. (1972). The location of rhythmic stress beats in English: An experimental study I. *Language and Speech*, **15**, 72–100.

Iwaidja Inyman Team. (2012). Ma! Iwaidja. https://itunes.apple.com/au/app/ma-iwaidja/id557824618?mt=8 (accessed 06.05.2014).

Kolly, M.-J. and A. Leemann. (in press). ‘Dialäkt Äpp’: Communicating dialectology to the public – crowdsourcing dialects from the public. To appear in A. Leemann, M.-J. Kolly, S. Schmid and V. Dellwo (Eds.), *Trends in Phonetics in German-speaking Europe*. Bern/Frankfurt: Peter Lang.

Leemann, A. and M.-J. Kolly. (2013). Dialäkt Äpp. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8 (accessed 06.05.2014).

Leemann, A., M.-J Kolly, and V. Dellwo. (2014). Crowdsourcing regional variation in speaking rate through an iOS app. To appear in: *Proceedings of Speech Prosody 2014*.

Künzel, H., H. Masthoff and Köster, J. (1995). The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. *Science and Justice*, **35**, 291–295.

de Vries, N., M. H. Davel, J. Badenhorst, W. D. Basson, F. de Wet, E. Barnard and A. de Waal. (2014). A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, **56**, 119–131.

Effect of the Double-Filtering effect on Automatic Voice Comparison

Jonas Lindh[1,] *and Joel Åkesson*[1]
[1]*Voxalys AB, Gothenburg, Sweden.*
{jonas|joel}@voxalys.se

In forensic casework today it is not uncommon to receive material recorded with mobile phones or other handheld recording devices. From experience we know most people do not treat recordings with as much care as a person well versed in audio technology. Especially given the varying circumstances under which the material can be recorded. Thus it is important we learn more about what sort of acoustic effects take place under particular conditions and how these effects can influence Automatic Voice Comparison (AVC). The current study aims at evaluating the effects of recording material consisting of what could be described as 'double-filtered' sound, henceforth referred to as DF, e.g. when a phone call is recorded using a handheld recorder placed in the vicinity of the mobile device. This filtering effect constitutes sound transmitted via GSM communication (1st filter) which then passes an indeterminable distance through the air before being captured by another recording device, such as a mobile phone or handheld recorder's microphone (2nd filter). This effect affects the energy in the signal. The energy decreases in both the low and the high frequencies, while the middle frequencies are boosted.

In this study we have used a database consisting of 150 female speakers of Swedish, all students of speech and language pathology. The recordings were made in a sound treated recording booth using a set-up of one computer equipped with an internal M-Audio soundcard and a high quality headset microphone. Each recording consists of solicited spontaneous speech together with read speech material (Swedish standard reading passage called 'Ett svårt fall'). Each speaker is informed and encouraged to finish the task at their own pace. Mean duration of the full recording among the speakers was 69.3 seconds (std 16 seconds).

Figure 1 Re-recording with double filtering in studio.

The DF effects have been evaluated using two AVC systems applying two different techniques, Batvox 4.1, (developed by Agnitio), a so called iVector system (Dehak et al.,

2009) and Vocalise (Oxford Wave Research) applying the so called UBM-GMM approach (Reynolds, 1992). Each recording in the database was split so that the read passage could be used as training material, while the spontaneous passage would be used for testing. For Batvox 100 speakers were used for testing, 50 speakers for score normalisation (30 speakers for T-norm and 20 speakers for Z-norm) (Barras and Gauvain, 2003). For Vocalise the same 100 speakers were used for testing and 50 speakers for the UBM.

The results show that normalisation techniques decreases the effect of the double filter.

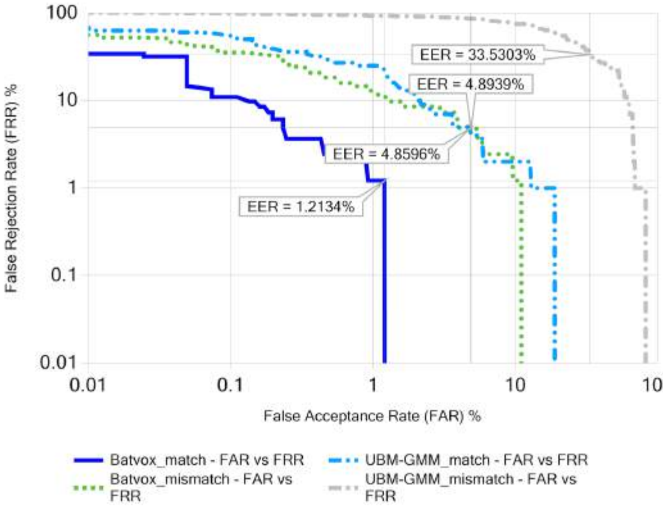


Figure 2 Log EER from the test results for both systems with both mismatched and matched training and test recordings.

In the next phase an error-check will be made to see whether the same mistakes are made by the two systems and between conditions. After that the material will be double-filtered using different recording distances to see how that affects the result.

References

Barras, C. and Gauvain, J.L. Feature and score normalization for speaker verification of cellular data, *Proc. of ICASSP*, April 2003.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P. and Dumouchel, P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in INTERSPEECH, Brighton, UK, Sept 2009.

Reynolds, D. A. A gaussian mixture modeling approach to text-independent speaker identification, Ph.D. thesis, Georgia Institute of Technology, August 1992.

Nativespeakerhood: A *Subject* Matter Revisited

Federal Office for Migration, Switzerland, LINGUA
lingua@bfm.admin.ch

Much of the discourse on nativespeakerhood in LADO has centered on whether non-linguist native speakers are more suitable to act as analysts than linguists who may not necessarily be native speakers (e.g. Cambier-Langeveld, 2010; Fraser, 2011). This paper, however, aims to shift the focus from the nativespeakerhood of the analyst to that of the asylum seeker and, in particular, its significance for actual LADO procedures.

While the language that asylum seekers specify as their native language plays a crucial role within the current asylum procedure in Switzerland, the value of this information towards determining the person's principal area of socialization varies from case to case. In an ideal scenario, an asylum seeker will claim to speak a language as his or her native language or 'mother tongue', which can then be confirmed or disproved by means of a linguistic analysis of their speech sample. In practice, however, these proceedings prove to be much less straightforward and raise a significant number of questions that need to be addressed both in general terms as well as with each individual biography. When an asylum seeker does not conform to what to an analyst sounds like a native speaker of a certain language, one needs to reassess the significance of the alleged 'native' language on multiple levels: On the one hand, it is necessary to consider sociolinguistic factors that prevail in the area in question, such as pidginization or existing basilect-acrolect continua as well as the consequences of diglossia or multilingualism. Which degrees of native-likeness can therefore be expected? On the other hand, in most cases it remains unclear what qualifies as 'native-like' in the first place. Does the linguistic evidence based on phonology, morpho-syntax and lexicon, for instance, suffice towards this end? And if not, what other measures can be taken?

In this paper, we draw on actual problems that arise among the current LADO practices of LINGUA for the asylum procedure in Switzerland. We aim to present real-world examples illustrating how the above issues are addressed as well as why and where further research on nativespeakerhood can prove useful for LADO.

References

Cambier-Langeveld, T. (2010) The role of linguists and native speakers in language analysis for the determination of speaker origin. *International Journal of Speech, Language and the Law*. *17*(01), 67-93. Retrieved April 15, 2014, from http://www.equinoxpub.com/journals/index.php/IJSLL/article/view/7318/6448.

Fraser, H. (2011). The role of linguists and native speakers in language analysis for the determination of speaker origin: A response to Tina Cambier-Langeveld. *International Journal of Speech, Language and the Law*. *18*(01), 121-130. Retrieved April 15, 2014, from http://www.equinoxpub.com/journals/index.php/IJSLL/article/view/9915/8621.

Listeners' perception of voice similarity in Standard Southern British English versus York English

Kirsty McDougall
Department of Theoretical and Applied Linguistics, University of Cambridge, UK
kem37@cam.ac.uk

This paper reports part of a research programme which explores 'perceived voice similarity' (PVS), the notion that within a group of speakers of the same sex and age, listeners will perceive certain speakers as sounding more similar to each other than others. Findings from this research focussing on various aspects of Standard Southern British English (SSBE) and on a voice parade case in East Anglia were presented at IAFPA 2010 and 2011 (McDougall 2011, Nolan *et al.* 2010, 2011). The present study considers the extent to which the acoustic correlates of PVS are consistent across accents, and reports results from a study of York English (YE) in comparison with the SSBE findings.

For the SSBE experiment, 15 male speakers, aged 18-25 years, were selected from the *DyViS* database (Nolan *et al.* 2009). For the YE experiment, 15 male speakers of the same age were selected from the newly developed *YorViS* database of YE[1] which contains recordings of the same format as *DyViS*. For both experiments, spontaneous speech from a telephone call task was used to create the stimuli: two utterances per speaker, each approximately 3 seconds in duration. Within each experiment, each speaker was matched with all other speakers and with himself to form 120 pairings. 20 listeners (speakers of British English; a different group for each experiment) rated the (dis)similarity of the paired voice samples on a nine-point scale from 'very similar' to 'very different'. Multidimensional scaling (MDS) was applied to the ratings to derive five perceptual dimensions for each accent whose correlation with long-term fundamental frequency, articulation rate, and long-term formant analysis of F1, F2, F3 and F4 was tested using Spearman's formula.[2]

Long-term fundamental frequency plays an important role in PVS in voice similarity, yielding significant correlations with perceptual dimensions in both accents. It correlates significantly with the first perceptual dimension for SSBE (= 0.804), indicating that it is of key importance for this accent. In YE, long-term f0 correlates significantly with dimension 3 (= 0.689), while the upper formants, F3 and F4, appear to show greater levels of importance by correlating significantly with respectively dimensions 1 (= 0.536) and 2 (= 0.514). F3 does not correlate with any perceptual dimension for the SSBE responses (at the time of writing, F4 information is not available for SSBE). The lower formants appear to behave in a similar fashion across the two accents, with F2 ranking higher (significant correlation with dimension 2 in SSBE (= 0.514) and with dimension 3 in YE (= 0.557)) than F1 (significant correlation with dimension 4 in both accents: SSBE = 0.675, YE = 0.718). Articulation rate did not achieve a significant correlation with any of the MDS dimensions in either accent, possibly due to the short duration of the stimuli.

Implications of the findings for voice parade construction, in particular with respect to the choice of foil voices, will be discussed.

References

K. McDougall (2011) 'Acoustic correlates of perceived voice similarity: a comparison of two accents of English.’ Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Vienna, 24-28 July 2011.

F. Nolan, J.P. French, K. McDougall, L. Stevens and T. Hudson (2011) ‘The role of voice quality ‘settings’ in perceived voice similarity.’ Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Vienna, 24-28 July 2011.

F. Nolan, K. McDougall, G. de Jong & T. Hudson (2009) The *DyViS* Database: Style-Controlled Recordings of 100 Homogeneous Speakers for Forensic Phonetic Research. *International Journal of Speech, Language and the Law,* **16.1**, 31–57.

F. Nolan, K. McDougall and T. Hudson (2010) ‘Perceived voice similarity and acoustic measures.’ Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Trier, 18-21 July 2010.

[1] British Academy Small Grant No. SG112892 to K. McDougall ‘A Cross-Accent Investigation of the Phonetic Correlates of Perceived Voice Similarity’.

[2] Thanks are due to Erica Gold, Philip Harrison and Christin Kirchhübel for their contributions to these phonetic analyses.

Eligible for student award

The Influence of Background Music on Perceived Speaker’s Age

Nancy Renning
Department of Comparative Linguistics, University of Zurich, Switzerland
nancy.renning@uzh.ch

The Influence of Background Music on Perceived Age of Speaker

When stimuli for phonetic experiments are chosen, background noises are usually minimized in favor of clear audio signals, which can more easily be compared against one another. However, research has so far largely neglected the question of whether subjects take background noises into account when interpreting audio stimuli. In this experiment, utterances in Swiss German drawn from the Dialäkt Äpp Corpus were combined with recordings of both classical and pop music.

Ninety subjects were asked in an online survey to estimate the age of a speaker in an audio stimulus. Subjects were able to complete the survey in about one minute, resulting in a high rate of return. In the analysis, the estimates that subjects made of speakers' ages while music played in the background were compared to those that other subjects made when the same stimuli were played without music or other sounds. The aim was to determine how background music affects the estimate itself, not the accuracy of that estimate. Furthermore, the actual age of the speakers, who submitted their data online via an app, was self-reported. This resulted in uncertainty as to whether the reported ages were the actual ages of the speakers.

It was found that pop music being played in the background led to subjects producing lower estimates of speaker' ages. The difference in estimates relative to the condition without music varied from 1 to 3.5 years. The effect was more pronounced when the age of the speaker was higher. However, the extent of this effect was significantly larger for some stimuli than others. In some cases, pop music did not have any significant influence. Classical music had no significant influence on age prediction across all stimuli.

Cognitive bias in forensic speech science

Richard Rhodes
J P French Associates & Department of Language & Linguistic Science, University of York, UK
richard.rhodes@jpfrench.com

Cognitive biases have been shown to have a detrimental effect on those forensic disciplines that rely on human interpretation (see Kassin, Dror & Kukucka, 2013, for a summary). The term *forensic confirmation bias* has been used to encompass a range of psychological processes that have the potential to affect judgements by forensic experts. These include exposure to inculpatory or strongly emotive contextual information, motivational factors (e.g. the desire to catch criminals - Charlton et al., 2010 - or find in favour of a client - Whitman & Koppl, 2010), primacy/order effects, expectancy effects related to frequency of positive outcomes and demographic stereotypes. The effects have been shown to be more damaging in cases where the data are incomplete or difficult to interpret (Dror, Charlton & Péron, 2006; Whitman & Koppl, 2010).

Although aspects of cognitive bias (chiefly priming) have been addressed in respect of forensic transcription/disputed content analysis by Fraser (2003; 2011), there has been relatively limited reflection on the potential for cognitive bias to affect forensic speaker comparison. This is particularly relevant for approaches which encompass subjective interpretation of results (i.e. those which do not rely on a numerical database for assessing strength of evidence).

There are a number of reasons why speech science might be more susceptible to these biases than other forensic disciplines. Unlike other forms of forensic science, such as DNA or toxicology, for example, analysts have a perceptual mechanism for speech and for recognising voices. They therefore might be more prone to early hypothesis-forming leading to the ‘tunnel vision’ described by Findley and Scott (2006). Moreover, unlike in fields such as DNA or toxicology, where the characteristics of the evidence are opaque to the instructing party, voice samples are likely to be pre-filtered and very different pairs/sets of voices filtered out. The similarity of samples and the incidence of positive results in speaker comparison, therefore, may well be greater than in other fields. Additionally, the prevalence (particularly in the UK) of using police interviews as reference material makes it more difficult to insulate analysts against potentially biasing contextual information about the case.

A number of recommendations for reducing the risk of cognitive bias have been made by psychologists, researchers and practitioners in other disciplines (Whitman & Koppl, 2010; Kassin, Dror & Kukucka, 2013). These include (but are not limited to):

- blind-testing (i.e. with no contextual information);
- testing within a line-up of suspect ‘foil’ samples;
- working in linear rather than cyclical fashion (from ‘crime’ to ‘known’ sample);
- verification by a second expert who is blind of the initial outcome;
- basic training relating to cognitive biases.

As a first step, this poster presentation will consider research concerned with reducing cognitive biases and bring it to bear on forensic speech science. I will be asking attendees and IAFPA members to fill in a questionnaire relating to bias in our field, the aim being to identify and share realistic and effective practices to manage bias.

References

Charlton, D., Fraser-Mackenzie, P. A., & Dror, I. E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis*. Journal of Forensic Sciences*, 55(2), 385-393.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications*. Forensic science international*, 156(1), 74-78.

Findley, K. A., & Scott, M. S. (2006). Multiple Dimensions of Tunnel Vision in Criminal Cases, *The. Wis. L. Rev*., 291.

Fraser, H. (2003) Issues in transcription: factors affecting the reliability of transcripts in legal cases. *International Journal of Speech, Language & the Law*, (10)2.

Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a crisis call: persistence of a primed perception of a disputed utterance. *International Journal of Speech, Language & the Law*, 19(2).

Kassin, S. M., Dror, I. E., & Kukucka, J (2013) The forensic confirmation bias: Problems, perspectives and proposed solutions. *Journal of Applied Research in Memory and Cognition* 2, 42-52.

Whitman, G., & Koppl, R. (2010). Rational bias in forensic science*. Law, probability and risk*, 9(1), 69-90.

Eligible for student award

Forensic voice comparison using glottal parameters in twins and non-twin siblings

Eugenia San Segundo Fernández[1], *Pedro Gómez-Vilda*[2]
[1]*Universidad Internacional Menéndez Pelayo (UIMP), Madrid, Spain*
eugeniasansegundo@gmail.com
[2]*Center for Biomedical Technology, Universidad Politécnica de Madrid, Spain.*
pedro@fi.upm.es

Forensic phoneticians have traditionally relied on the information found in the vocal folds for speaker identification: from the analysis of classical distortion parameters like jitter and shimmer (Künzel & Köster, 1992) and other laryngeal features (Jessen, 1997) to the automatic approaches exploring the usefulness of the combined use of vocal source and vocal tract information in order to improve speaker-recognition systems (Zheng, 2005; Farrús, 2008). Based on previous voice-pathology investigations (Gómez et al., 2007), other studies by these authors have more recently shown that their voice-analysis methodology based in the decoupling of vocal tract from glottal source estimates can also be useful for the biometric characterization of speakers (Gómez et al., 2009, 2010). Following these studies, San Segundo (2012) was a pilot experiment with a relatively small sample of MZ and DZ twins (12 and 8, respectively), and using only some of the glottal parameters provided by a specific software package implementing vocal tract inversion and glottal source parameterization (www.biometrosoft.com).

For the current study 54 speakers were recruited: 24 MZ pairs, 10 DZ pairs, 8 non-twin brothers and 12 reference-population speakers. In a first step, as a follow-up of San Segundo (2012), some reexaminations and in-depth voice analyses were carried out for all the 20 speakers' voices already analyzed in the above-mentioned proof of concept: 1) anamnesis reexamination to discard possible voice-related pathologies; 2) reexamination of the parameter values extracted, since the analysis in the pilot experiment was carried out in a batch-mode and this kind of processing may entail certain evaluation software artifacts (ESA); and 3) new voice analysis and back annotation with the aim of visually inspecting the glottal waveform of the speakers' voices and checking their fitting to usual normophonic thresholds. Besides, if deemed necessary, a DNA test was carried out to confirm the twins' zygosity. In a later step, the naturally-sustained [e:] fillers of all the 54 speakers (2 sessions per speaker) were extracted and analyzed with the same software creating a vector of 68 parameters from each vowel segment, comprising: 1) f0 and distortion parameters; 2) cepstral coefficients of the glottal source power spectral density (PSD); 3) singularities of the glottal source PSD; 4) biomechanical estimates of vocal fold mass, tension and losses; 5) time-based glottal source coefficients; 6) glottal gap (closure) coefficients; and 7) tremor (cyclic) coefficients. Finally, a forensic comparison was carried out using the methodology described in Gómez et al. (2012). The results suggest that the parameters analyzed are somehow genetically related, as more similarity is found in MZ twins than in DZ twins or non-twin siblings. Besides, the between-speaker comparisons for unrelated speakers yield LLRs homogeneously around -10, indicating a very good performance of the system.

References

Farrús, M. (2008). Fusing prosodic and acoustic information for speaker recognition. PhD Thesis, Universitat Politècnica de Catalunya.

Gómez, P., Fernández-Baillo, R., Nieto, A., Díaz, F., Fernández-Camacho, F.J., Rodellar, V., Alvarez, A. and R. Martínez. (2007). Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. *Journal of Voice*, **21** (4), 450-476.

Gómez, P., Fernández-Baillo, R., Rodellar, V., Nieto V., Álvarez-Marquina, A., Mazaira-Fernández, L.M., Martínez, R. and J.I. Godino. (2009). Glottal Source biometrical signature for voice pathology detection. Speech Communication, **51** (9), 759-781.

Gómez, P., Álvarez-Marquina, A., Mazaira-Fernández, L.M., Fernández-Baillo, R., Rodellar, V. and V. Nieto. (2010). Glottal Biometric Features: Are Pathological Voice Studies appliable to Voice Biometry? *WTM-IP Workshop Tecnologías Multibiométricas para la Identificación de Personas*, Universidad de Las Palmas de Gran Canaria, 4-6 July 2010.

Gómez, P., Mazaira-Fernández, L.M., Martínez, R., Álvarez-Marquina, A., Hierro, J.A., and R. Nieto. (2012). Distance Metric in Forensic Voice Evidence Evaluation using Dysphonia-relevant Features. *VI Jornadas de Reconocimiento Biométrico de Personas (JRBP)*, Las Palmas de Gran Canaria, 26-27 January 2012.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *Forensic Linguistics* **4**(1): 84-103.

Künzel, H.J. and J-P. Köster. (1992). Measuring Vocal Jitter in Forensic Speaker Recognition. *Proceedings 44th Annual Meeting American Academy of Forensic Sciences*, New Orleans, 113-114.

San Segundo, E. (2012). Glottal source parameters for forensic voice comparison: An approach to voice quality in twins' voices, *21th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, Santander, 6-8 August 2012.

Zheng, N. (2005). Speaker recognition using complementary information form vocal source and vocal tract. PhD Thesis, The Chinese University of Hong Kong.

Eligible for student award

Perceptual speaker discrimination based on German consonants

Carola Schindler, Eva Reinisch, Jonathan Harrington
Institute of Phonetics and Speech Processing, Ludwig Maximilians University, Munich, Germany
{carola.schindler|evarei|jmh}@phonetik.uni-muenchen.de

Nasal and fricative consonants appear to contain high amounts of speaker-specific information. Their acoustic properties tend to differ between speakers to a larger extent than within one speaker's different productions. Mook and Draxler (2012) showed this for German by analysing spectral moments of different types of consonants and vowels (see also Schindler & Draxler, 2013). Speaker-specific information in nasals and fricatives is also evident in perception. In a speaker discrimination experiment participants were more accurate when the words they heard contained /m/ or /s/ than /l/ or /t/ (Andics, 2013). The goal of the present study was to explore in more detail and with a larger set of consonants whether listeners' ability to perceptually discriminate between speakers depends on the types of consonants they hear and whether this pattern would match the acoustic analyses.

Participants performed a speaker discrimination task. Stimuli were nonsense words that consisted of a consonant in a bilateral /a/-context. Consonants were nasals, fricatives, and stops, each in labial and alveolar place of articulation (i.e., /m/, /n/, /f/, /s/, /p/, /t/). Four different tokens from nine Bavarian speakers were recorded and paired to same-speaker and different-speaker pairs. Participants performed a same-different discrimination task. Each stimulus was flanked by 500 ms pink noise to make the task harder. Trials were blocked by consonant and presented in randomised order.

Since listeners were very good at discriminating speakers for all consonants (mean accuracy in a pilot study was 0.95) and in order to reach a better degree of separation between consonants, the stimuli were manipulated to make the task more difficult. The pitch contour was flattened and normalised, and the vowels shortened to 50 ms on each side. This caused the overall accuracy to drop to 0.83. In a second experiment the consonants were spliced into an identical vowel context for all speakers, so that the listeners could not use the vowel information to discriminate between the speakers (mean accuracy 0.62). Both experiments showed differences between the consonants, with larger effects for the more difficult task. Also the place and manner of articulation modulated listeners' speaker-discrimination abilities. Comparing accuracy rates for the different types of manipulated stimuli (with or without speaker information in the vowel) will also help to pinpoint what kinds of information contribute to speaker discriminability.

References

Andics, A. (2013). Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. PhD Thesis, Radboud University Nijmegen, Nijmegen.

Mook, C. and Draxler, C. (2012) The speaker discriminating power of nasals, fricatives and vowels, Poster presented at the IAFPA 2012, Santander, Spain.

Schindler, C. and Draxler, C. (2013) The influence of the place of articulation on the speaker specificity of German phonemes, Poster presented at the 4th International Summer School on "Speech Production and Perception: Speaker-Specific Behavior" 2013, Aix-en-Provence, France.

Earwitness speaker identification and physiological responses

Maartje Schreuder[1,2]*, and Thomas Meyer*[1]

[1]*Department of Psychology and Neuroscience, Maastricht University;*
[2]*The Maastricht Forensic Institute, The Netherlands*
Maartje.Schreuder@maastrichtuniversity.nl
Thomas.Meyer@maastrichtuniversity.nl

The current study was inspired by a case in which a robbery victim, long after the robbery, had strong physiological reactions and felt reminded of the crime after hearing a certain voice. However, the victim did not actually recognize this voice as the perpetrator's. For the police, the question emerged whether bodily reactions to a voice are a sign of implicit voice recognition, and how accurate this may be as compared with explicit voice recognition. To date, no experimental data are available to provide a direct answer to this question. Therefore, our fist aim was to test whether hearing the voice of a perpetrator can trigger emotional memories and enhance startle reactions in earwitnesses. Our second aim was to explore whether enhanced startle responses to hearing a perpetrator's voice go hand in hand with better voice recognition.

We exposed 84 healthy participants to an emotional audio clip of a staged bus hijack. After a 30-minute retention interval, they underwent an earwitness identification paradigm that was combined with a startle paradigm (Meyer et al., in press) to measure how the memories associated with the presented voices modulate affective responses. In particular, participants heard 84 neutral and negative voice fragments spoken by the (acting) bus hijacker and two foil speakers, and indicated whether they recognized the voice as the perpetrator's. The trials were accompanied by light flashes to elicit eye-blink startle reflexes (Bradley & Lang, 2000). In this presentation, we will present the results on implicit speaker recognition in terms of startle modulation, and its relation to explicit speaker recognition.

References

Bradley, M.M. and P.J. Lang (2000). Affective reactions to acoustic stimuli. *Psychophysiology,* **37***,* 204–215.

Meyer, T., Quaedflieg, C.W.E.M., Giesbrecht, T., Meijer, E., Abiad, S., and T. Smeets (in press). Frontal EEG asymmetry as predictor of physiological responses to aversive memories. *Psychophysiology*.

Comparison of similarity and dissimilarity indices between speech samples in filtered and non-filtered conditions for the speakers of the Croatian language

Gordana Varošanec-Škarić[1], *Iva Pavić*[2], *and Gabrijela Kišiček*[2]
[1, 2, 3]*Department of Phonetics, University of Zagreb, Croatia*
gvarosan@ffzg.hr, ipavic2@ffzg.hr, gkisicek@ffzg.hr

Acoustic-statistical measurements of similarity index (R) and dissimilarity index (SDDD) on the basis of long term average spectra (LTASS) can be used as a support measurement in forensic phonetic cases (Harmegnies, 1995). In this research similarity and dissimilarity indices were compared for speech samples in filtered and non-filtered conditions. The data consisted of 86 speakers originating from 8 largest Croatian cities representing three dialects of Croatian language. All speakers were recorded under the controlled conditions reading standardized text and during the spontaneous speech. Recordings were edited in Cool Edit program and speech samples (duration 60 s) were filtered. Filtered and non-filtered speech samples were than compared on the basis of LTASS (non-filtered conditions (0 - 10 kHz) and filtered (0.8 – 4 kHz)). Using index R and index SDDD intraspeaker variations and interspeaker variations were compared respectively for male and female speakers. Results of intraspeaker variations showed that average values of similarity index (R) in non-filtered conditions were between 0.94 for male speakers in reading texts to 0.98 for female speakers in reading text and spontaneous speech. Results of interspeaker variations showed lower values of index R in the non-filtered conditions: from 0.86 in spontaneous speech to 0.94 in reading text for female speakers. Average values of R in filtered conditions for intraspeaker variations were between 0.83 for both female and male speakers in spontaneous speech to 0.95 in reading texts. Average values of R index in filtered conditions for interspeaker variations were significantly lower; from 0.57 for male spontaneous speech to 0.9 for female reading texts. Average values of index SDDD in non-filtered conditions for intraspeaker variations were generally lower – from 2.27 for female speakers to 3 for male speakers in reading. SDDD index showed higher values in non-filtered conditions for interspeaker variations; from 4.75 in female reading speech and male spontaneous speech to 5.12 for male reading speech. In filtered conditions intraspeaker variations resulted with SDDD index between 2.14 for male reading speech to 3.01 for female spontaneous speech. As expected, results in filtered conditions for interspeaker variations showed higher values of SDDD index, from 3.06 for female to 4.71 for male reading speech. The differences between similarity index (R) in intraspeaker variations were statistically significant for female speakers ($p<0.0001$) and for male speakers ($p<0.05$) in both spontaneous speech and reading. Results of interspeaker variations showed statistically significant differences in similarity index (R) for male speakers ($p<0.0001$ in reading and $p<0.0001$ in spontaneous speech) and female speakers ($p<0.0001$ in reading and $p<0.0001$ in spontaneous speech) and statistically significant dissimilarity index (SDDD) differences for male speakers ($p<0.0001$ in reading and $p<0.0001$ in spontaneous speech) and female speakers ($p<0.0001$ in reading and $p<0.0001$ in spontaneous speech). Overall results of this research show that acoustic-statistical measurement of similarity and dissimilarity indices are a useful method in speaker recognition in forensic phonetic expertise. Further on, results show that speaking conditions should not be neglected in forensic phonetic cases.

References

Baldwin, John R., Peter French (1990). *Forensic Phonetics.* London and New York: Pinter Publishers.

Boersma, Paul, David Weenink (2009). Praat: doing phonetics by computer, version 5.1.20 www.fon.hum.uva.nl/praat/ (1. prosinca 2013.)

French, Peter (2013). Forensic speaker comparison: man and machine – Forenzična usporedba govornika: čovjek i stroj. U: Vlašić Duić, J. i Varošanec-Škarić, G. (ur.) *Knjiga sažetaka – Istraživanja govora*, Zagreb: Filozofski fakultet Sveučilišta u Zagrebu, Hrvatsko filološko društvo. Osmi znanstveni skup s međunarodnim sudjelovanjem, od 5. do 7. prosinca 2013., str. 25-26.

Harmegnies, Bernard (1995). Contribution a la caracterisation acoustique des sigmatismes – etude de deux indices acoustico-statistiques. U: A. Braun i J.-P. Köster (ur.) *Studies in Forensic Phonetics,* 56-66. Trier: Wissenschaftlicher Verlag Trier.

Harmegnies, Bernard, Albert Landercy (1985). Language Features in the Long-Term Average Spectrum. *Revue de Phonétique Appliquée* 73-74-75, 69-79.

Harrison, Philip, Peter French (2010). Assessing the suitability of BATVOX for UK Casework or Evaluation of the BATVOX automatic speaker recognition system for use in UK based forensic speaker comparison casework Part II. U *Abstracts for the 19th Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Trier, Germany, Department of Phonetics, University of Trier, str. 13.

Hollien, Harry (2002). *Forensic Voice Identification.* San Diego: Academic Press.

Künzel, Hermann J. (2010). Automatic Speaker Identification with Multilingual Speech Material. U *Abstracts for the 19th Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Trier, Germany, Department of Phonetics, University of Trier, str. 20.

Künzel, Hermann J. (2013). Automatic speaker recognition with cross-language speech material. *Journal of Speech, Language and the Law* 20.1, 21-44.

Labov, William (1972). *Sociolinguistic Patterns.* Philadelphia: Philadelphia University of Pennsylvania Press.

Labov, William (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics* 34, 500-515.

Nolan, Francis (1983, digitally printed version 2009). *The phonetic bases of speaker recognition*. CambridgeCambridge: University Press.

Nolan, Francis (2007). Voice quality and forensic speaker identification. *Govor* **XXIV**, 2, 111-128.

Nolan, Francis, Catalin Grigoras (2005). A case for formant analysis in forensic speaker identification. *Speech, Language and the Law* 12, 2, 143-173.

Rodman, Robert, David F. McAllister, Donald L. Bitzer, Luis F. Cepeda, Pamela Abbitt (2002). Forensic speaker identification based on spectral moments. *Forensic Linguistics* 9, 1, 22-43.

Rose, Phil (2002). *Forensic Speaker Identification.* London, New York: Taylor and Francis.

Varošanec-Škarić, Gordana (2008). Speaker verification in Forensic Phonetics. *Govor – časopis za fonetiku* XXV, 1, 31-44.

Varošanec-Škarić, Gordana, Jordan Bićanić (2007). A comparison of indices of difference and similarity based on voices in real forensic case and in controlled conditions. *Proceedings*. www.icphs20007.de, 16th International Congress of Phonetic Sciences (Eds. Jürgen Trouvain i William J. Barry), pp 2085-2088. (3. prosinca 2013.)

Varošanec-Škarić, Gordana, Gabrijela Kišiček (2012). Forensic Phonetic identification and linguistic analysis of the speaker. *Suvremena lingvistika* 73, 89-108.

Wolfram, Walt, Ralph W. Fashold (1997). Field methods in the study of social dialects. U: Coupland N., Jaworski A. (ur.) *A Sociolinguistcs. A Reader and a Coursebook.* London: Macmillian, 89-115.

Eligible for student award

Stability of short-term voice quality parameters in GSM

Jitka Vaňková, Tomáš Bořil, and Radek Skarnitzl
Institute of Phonetics, Faculty of Arts, Charles University in Prague, Czech Republic
jitka.vanka@gmail.com, {tomas.boril|radek.skarnitzl}@ff.cuni.cz

Voice quality parameters have not been investigated to a great extent in technical speaker identification tasks, in spite of the fact that forensic phoneticians appear to make rather frequent use of voice quality in their casework (Nolan, 2005; Gold & French, 2011). The main reason for the lack of acoustic investigations – one of the few exceptions being Jessen (1997) – appears to be the fact that the presence of especially laryngeal voice quality features is compromised in telephone speech (Nolan, 2005). In addition, the plasticity of our voice production mechanism allows for great stylistically conditioned variability, and it is mostly voice quality which is affected.

Yet we believe that there still is space for acoustic examinations of speaker specificity of voice quality, especially of its short-term correlates which reflect spectral slope by comparing the amplitudes of various events in the acoustic spectrum (Hanson et al., 2001). The motivation for using the parameters H1*-H2*, H1*-A1*, H1*-A2*, H1*-A3* and H2*-H4* is twofold: first, it appears that some of them yield favourable rates of intra-speaker stability and inter-speaker variability (Vaňková and Skarnitzl, 2014); second, low frequencies relevant for H1 are actually not filtered out by the Adaptive Multi-Rate (AMR) codec, which is the current standard in mobile telephony (Guillemin and Watson, 2008; Vaňková and Bořil, submitted).

We analyzed recordings of 5 female and 5 male speakers which were passed through the AMR codec, using the lowest and highest bit rate of both its narrowband (NB) and wideband (WB) version (3GPP, 2012). 15 vowel items of each of the short Czech monophthongs /ɪ ɛ a o u/ were used, yielding 750 vowel tokens. F0 and formants were extracted from the central part of each token, and voice quality parameters computed using VoiceSauce (Shue, 2013) five times – from the original studio recordings and from the four types of GSM compression.

Table 1 displays mean differences between parameter values (in dB) in studio recordings and the four codec conditions (positive values signal higher studio values, negative ones higher codec values). Differences in values depend on individual parameters (H1*-A2* and H2*-H4* appearing most robust; note that H1*-H2* was identified by Jessen, 1997 as carrying the most speaker-specific information) and they also vary across codec conditions (WB performing overall better than NB). In terms of other sources of variability, the impact of the codec on the parameters was likewise found to differ for the two genders and individual vowel qualities. These results will also be included in the presentation.

Table 1. Mean differences between parameter values (in dB) in studio recordings and the four codec conditions. Positive values signal higher studio values; negative higher codec values.

	studio-NB 4.75	***studio-NB 12.20***	***studio-WB 6.60***	***studio-WB 19.85***	**Total (abs)**
H1*-H2*	1.46	0.70	0.97	0.19	**3.32**
H2*-H4*	0.55	0.13	0.59	0.43	**1.70**
H1*-A1*	2.52	1.35	1.86	0.64	**6.37**
H1*-A2*	-1.03	-0.27	-0.17	0.17	**1.64**
H1*-A3*	-1.76	-1.21	-0.43	0.42	**3.82**
Total (abs)	**7.33**	**3.65**	**4.02**	**1.84**	

References

3GPP (2012). TS 26.071 AMR speech CODEC; General description. Downloaded on February 2, 2014 from http://www.3gpp.org/ftp/specs/-archive/26_series/26.071/.

Gold, E. and P French. (2011). International practices in forensic speaker comparison. *The International Journal of Speech, Language and the Law,* **18**, 293–307.

Guillemin, B. J. and C Watson (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language and the Law*, **15**, 193–218.

Hanson, H. M., K N Stevens, H-K J Kuo, M Y Chen and J Slifka (2001). Towards models of phonation. *Journal of Phonetics*, **29**, 451–480.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *Forensic Linguistics*, **4**, 84–103.

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality". In W. Hardcastle & J Beck (Eds,), *A Figure of Speech*, 385–411. Mahwah, New Jersey: Erlbaum.

Shue, Y. (2013). VoiceSauce: A program for voice analysis (V1.14). Downloaded on October 7, 2013 from http://www.seas.ucla.edu/spapl/voicesauce/

Vaňková, J. and T Bořil (submitted). Impact of the GSM AMR codec on automatic vowel formant measurement in Praat and VoiceSauce (Snack). EUSIPCO 2014.

Vaňková, J. and R Skarnitzl (2014). Within- and between-speaker variability of parameters expressing short-term voice quality. *Proceedings of Speech Prosody 2014*, 1081–1085. Dublin.

NFI-FRITS: A forensic speaker recognition database

David van der Vloed[1], *Jos Bouten*[2]
[1]*Netherlands Forensic Institute*
d.van.der.vloed@nfi.minvenj.nl
[2]*visiting scientist*

The NFI-FRITS database (Forensically Realistic Intercepted Telephone Speech) contains speech intercepted during real police investigations. This material was obtained to facilitate research on data typically encountered in forensic practice, much like the data used by Becker (2012) and Van Leeuwen and Bouten (2004) and the AHUMADA III data (Ramos et al, 2008). NFI-FRITS consists of over 4100 recordings of more than 600 speakers.

Data processing

The raw data were provided with some metadata, like the two telephone numbers involved in the telephone call, case name, etc. The audio files were split in two single channel files (a and b) and stored in a database, along with the provided metadata.
Native listeners listened to the material and removed information in the audiofiles that can identify an individual and assigned speaker names and other metadata. This was done until about five recordings were assigned to a speaker, after which the process was repeated.

Realistic data

The database consists of realistic data, meaning that the audio comes from intercepted telephone speech from police investigations. The forensic nature of the recordings and the method to label a recording with a speaker name make the truth about speaker identities in this database a truth by proxy. Nevertheless the authors feel that the method used leads to sufficiently reliable speaker identities. The data is representative of police investigations, however, the collection is not representative of casework at the NFI as this typically involves recordings where the speaker ID is disputed.

Database by numbers

The database consists of 4188 recordings and 604 speakers. There are 427 male speakers in 3120 recordings and 177 female speakers in 1068 recordings. There are 72 multilingual speakers in the database, who speak Dutch in some recordings and either Turkish, Moroccan Arabic or Berber in other recordings.

Table 1. No. recordings per language

Language	#recordings
Dutch (all varieties)	3091
Turkish	499
Moroccan Arabic	191
Berber (Tarifit)	116
Mixed	245
Other	46

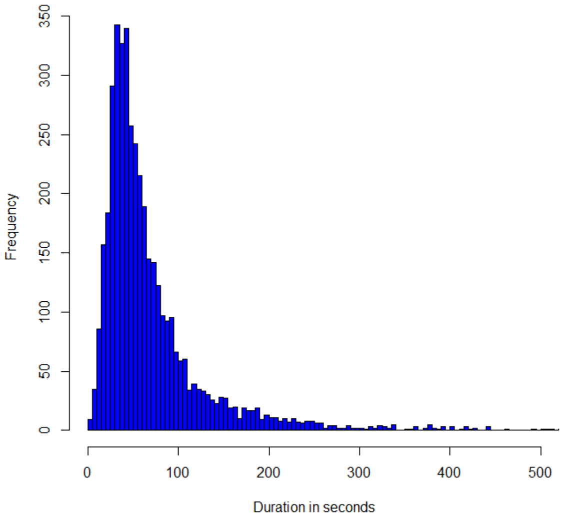


Figure 1 Histogram of measured durations of speech per recording

References

T. Becker, “Automatic forensic voice comparison (automatischer forensischer stimmenvergleich),” in The Journal of Speech Language and the Law, 2012, vol. 19, pp. 291–294.

David A. van Leeuwen and Jos S. Bouten, “Results of the 2003 NFI-TNO forensic speaker recognition evaluation,” in Proc. Odyssey 2004 Speaker and Language recognition workshop. June 2004, pp. 75–82, ISCA.

Daniel Ramos, Joaquin Gonzalez-Rodriguez, Javier Gonzalez-Dominguez, and Jose Juan Lucena-Molina, “Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish.,” in Proc. Interspeech, 2008, pp. 1493–1496.

Ratings of ‘threat’ and ‘intent’ by listeners exposed to neutrally-worded utterances in five languages

Dominic Watt, Sarah Kelly, and Carmen Llamas
Department of Language & Linguistic Science, University of York, UK
{dominic.watt|sk720|carmen.llamas}@york.ac.uk

Research on indirect threats – speech acts also interpretable as, say, warnings, advice or neutral statements of fact/opinion, e.g. *I wouldn’t go talking to the police,* or *It’d be a shame if something were to happen to your kids* – has tended to focus on the linguistic content and the context-dependency of such utterances, and the extent to which readers/hearers interpret them as threats on the basis of these two factors (Fraser, 1998, Gales, 2012). Legally speaking, a threat only becomes one when it is treated as such by an observer who, on the basis of spoken or written words, forms beliefs about the intention and the capacity of the person delivering the threat to cause harm to the recipient and/or to a third party. The UK Public Order Act (1986, Ch. 64/4.1) specifies that use of threatening words is an offence if the hearer thereby has reason to believe that the speaker intends to perform an act that would be harmful to the hearer or to other individual(s).

We report an experiment seeking to elicit listeners’ subjective ratings of neutrally-worded utterances designed to convey threat, and identically-worded ones which were not. The sentence *I know where you live* was read aloud by 8 adult male British English speakers, in 4 conditions, A-D. In condition A, speakers read the sentence, presented in isolation with no prompting of any sort, using a ‘normal’ tone of voice. In the ‘induced-threat’ condition B, speakers were asked to read the sentence in a ‘threatening’ way, where this was to be interpreted as they liked. In condition C, speakers read a short script incorporating the target sentence. The wording made it clear that the text was meant to be read in a non-threatening manner. Finally, condition D (‘induced-threat’) was also scripted, but on this occasion the message was clearly intended to encode an attempt to intimidate the recipient. The target sentence was then extracted from the C and D recordings, to remove contextual cues.

So as to test whether the wording itself was perceived to contribute to the perception of threat, the test sentence and accompanying scripts were translated into 4 other languages (Arabic, Swedish, Hebrew, Norwegian), and read aloud by native speakers of those languages. A panel of native English-speaking listeners (N=30), screened for knowledge of any of the foreign languages, were asked to rate the randomised English and foreign-language sentences for perceived threat level and for ‘intent to harm’. We predicted that these parameters would be closely correlated, but – there being such things as ‘empty threats’ – we thought it important to elicit judgements about both separately.

Listeners could distinguish induced-threat utterances from neutral ones, but did so more consistently for the English utterances than they did for the foreign-language ones. An understanding of the linguistic content of the utterance clearly allows listeners more readily to interpret (simulated) indirect threats as such; on its own, ‘tone of voice’ appears to have a relatively minor effect, albeit a potentially pivotal one. We conclude by outlining a planned series of experiments that will give us further insights into this hitherto unexplored area of forensic speech science.

References

Fraser, B. (1998). Threatening revisited. *Int. J. of Speech, Language & the Law* **5(2)**, 159–173.

Gales, T. (2012). Linguistic analysis of disputed meanings: Threats. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell. Online resource. DOI: 10.1002/9781405198431.wbeal0711

Scherer, K. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech and Language* **27**, 40–58.

Eligible for student award

Filled pauses as variables in speaker comparison: dynamic formant analysis and duration measurements improve performance for *um*

Sophie Wood, Vincent Hughes, and Paul Foulkes

Department of Language and Linguistic Science, University of York, York, UK.
{sophie.wood|vh503|paul.foulkes}@york.ac.uk

It is often hypothesised that filled pauses (FPs, i.e. *uh*, *um*) are useful variables in forensic speaker comparison (e.g. Künzel 1997, Tschäpe et al. 2005, Foulkes et al 2004, Jessen 2008). They offer several potential advantages over traditional segmental variables:

1. they are very frequent for most speakers and in most types of spontaneous speech;
2. they are typically longer than lexical vowels, and generally easier to measure;
3. they often abut silence, rendering them less susceptible to coarticulation, and thus in principle more consistent for the individual speaker;
4. there may be idiosyncratic patterns in the overall frequency of use, and in the discourse or syntactic contexts in which hesitations are used;
5. f0 patterns and durations may vary, as well as spectral components of vocalic elements;
6. the relative proportions of different FP types may also vary across speakers, i.e. whether speakers use vowel only (*uh*) or vowel+nasal (*um*) markers.

Here we present a study to investigate the discriminatory power of FPs, extending preliminary work presented by King et al (2013). FPs for 75 young male speakers of standard British English were analysed, drawn from Task 1 of the DyVis corpus (Nolan et al. 2009). The following acoustic properties were examined: 'static' midpoint frequencies of the first three formants in the vocalic portion; 'dynamic' measurements of the formants (i.e. quadratic curves fitted to 9 measurement points over the full vowel); and duration. Contemporaneous likelihood ratios were computed for independent sets of 25 development and 25 test speakers in MatLab (Morrison 2007) using Aitken & Lucy's (2004) Multivariate Kernel Density (MVKD) formula. Typicality was assessed using a reference set consisting of 25 speakers. Calibration coefficients were calculated based on the scores from the development data using a robust implementation of Brümmer's (2007) logistic regression procedure (Morrison 2009). The coefficients were then applied to the scores from the test data to generate calibrated log LRs. System performance was assessed using (i) Equal Error Rate (EER) as a metric of absolute discrimination between SS and DS pairs, and (ii) the log LR cost function (C_{llr}) (Brümmer & du Preez 2006), which provides a gradient assessment of system accuracy based on the magnitude of contrary-to-fact LRs.

Results are summarised in Table 1. For *uh* the static measurements outperform the dynamic measurements: EER is the same or slightly worse with the dynamic measurements, and C_{llr} is markedly worse in the dynamic measurement tests. This may be due to issues of overfitting trajectories that are essentially flat throughout the *uh* vocoid, meaning that static midpoints provide as much information without requiring so much input data. For *um,* on the other hand, dynamic measurements perform better than static measurements: EERs fall to less than 5% and C_{llr} reduces to less than 0.2. It is likely that the dynamic properties of *um* are more useful than those for *uh* because /VN/ FPs contain inherently more acoustic change between the vocalic and nasal portions. The addition of duration information further improves the EER and C_{llr} for *um.*

This study obtains LRs with EER scores below 5% using acoustic-phonetic features in spontaneous speech recordings, which compares well with studies such as Becker, Jessen and Grigoras (2008). The study therefore strongly supports the view that FPs have excellent potential as variables in forensic speaker comparison cases, although formant dynamic data may only be useful for *um*, whereas static measurements provide equally good or better results for *uh*.

Table 1. Summary of results for *uh* and *um*.

Test:		EER (%):	Cllr:
Static	*Uh*	11.92	0.5246
	Um	11.92	0.3692
Static + duration (Static measurements fused with durations)	*Uh*	12.00	0.4876
	Um	8.92	0.2825
Dynamics	*Uh*	15.17	0.7068
	Um	4.67	0.1978
Dynamics + duration (Dynamic measurements fused with durations)	*Uh*	11.92	0.7449
	Um	4.17	0.1821

References

Aitken, C.G.G. & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics,* **54**, 109-122.

Becker, T., Jessen, M. & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian Mixed Models. Paper presented at ISCA conference, Brisbane, Australia.

Brümmer, N. & du Preez, J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language,* **20**, 230-275.

Foulkes, P., Carrol, G. & Hughes, S. (2004). *Sociolinguistics and acoustic variability in filled pauses.* Paper presented at IAFPA conference, Helsinki, Finland.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass,* **2**, 671-711.

King, J., Foulkes, P., French, P. & Hughes, V. (2013). Hesitation markers as a parameter for forensic speaker comparison. Paper presented at IAFPA conference, Tampa, Florida, USA.

Künzel, H.J. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics,* **4**, 48-83.

Nolan, F., McDougall, K., de Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *The International Journal of Speech, Language and the Law*, **16**, 31-57.

Tschäpe, N., Trouvain, J., Bauer, D. & Jessen, M. (2005). *Idiosyncratic patterns of filled pauses.* Paper presented at IAFPA conference, Marrakesh, Morocco.

Eligible for student award

Speaker Profiling: An automatic method?

Georgina Brown, and Jessica Wormald
Department of Language and Linguistic Science, University of York, York, UK
{gab514|jessica.wormald}@york.ac.uk

Increasing attention is being given to the application of automatic speaker systems in forensic casework. The current paper considers an automatic speaker profiling system developed using the ACCDIST metric (Huckvale 2004) to group speakers into accent groups (Huckvale, 2007; Ferragne and Pellegrino, 2007; Hanani *et al*, 2013). The current system allows for the clustering of phones, meaning phoneme categories can be compared, not just individual, context dependent, segments. The system relies on segmental input in the form of mid-point MFCC vectors. Potential features of interest can both be specified, and also potentially identified through the system. This paper investigates the effects of various segmental combinations, exploring the system's strengths and weaknesses in a classification task. It highlights the importance of making considered phonemic choices when training the system before generating an automatic result. The first author has previously used the modified system with speakers of Scottish/English border varieties and observed a 61.2% recognition rate on an eight way recognition task with speakers from four locations and two age groups (Brown & Watt 2014).

The current paper demonstrates the system’s ability to classify groups of Panjabi-English (PE) speakers across and within the two English cities of Bradford and Leicester after training with reading passage data. PE speakers are British-born native-English speakers with Panjabi language heritage. Within each location, two age groups are represented.

Results considering 20 PE speakers from Bradford and 26 from Leicester highlight the ability of the system to recognise speakers from different geographical locations and of different ages. Table 1 includes the system’s results when including all vowel phonemes. Variation from these results is observed depending upon the combination of features selected, highlighting the importance of grounded sociophonetic choices when training the system. Sociophonetic differences between the respective groups can be exploited to improve the system’s results

Table 1. Results from the Phoneme-based ACCDIST system. Results for all vowels and best reduced combination included.

PE Speaker Groups	*Features*	N correct	% correct
Bfd old vs. Lei old	All vowels	15/18	83.3
	FACE GOAT PRICE MOUTH	18/18	100
Bfd young vs. Lei young	All vowels	25/26	91.2
	FACE GOAT PRICE MOUTH	23/26	88.5
Bfd old vs. Bfd young	All vowels	5/20	25.0
	FACE GOAT PRICE MOUTH + /r/	15/20	75.0
	CHOICE NEAR FLEECE + /r/	13/20	65.0
Lei old vs. Lei young	All vowels	13/24	54.2
	FLEECE KIT GOOSE FOOT	17/24	70.8

References

Brown, G. and D. Watt (2014). *Performance of a novel automatic accent classifier system using geographically-proximate accents*. Poster presented at BAAP, University of Oxford. 7th-9th April

Ferragne, E. and F Pellegrino. (2007). Automatic dialect identification: A study of British English. In Christian Muller (Ed.), *Speaker Classification, Volume 2 of Lecture Notes in Computer Science*. Berlin Heidelberg: Springer-Verlag. pp 243-257.

Hanani, A., M. Russell and M. Carey. (2013). Human and computer recognition of regional accents and ethnic groups for British English speech. *Computer Speech and Language*, **27**, 59-74.

Huckvale, M. (2004). ACCDIST: A metric for comparing speakers' accents. In *Proceedings of the International Conference on Spoken Language Processing.* pp. 29-32. Korea.

Huckvale, M. (2007). ACCDIST: An accent similarity metric for accent recognition and diagnosis. In Christian M̈üller (Ed.), *Speaker Classification, Volume 2 of Lecture Notes in Computer Science*. Berlin Heidelberg: Springer-Verlag. pp. 258-274.

NFI-FRITS: A forensic speaker recognition database

David van der Vloed[1], *Jos Bouten*[2]
[1]*Netherlands Forensic Institute*
d.van.der.vloed@nfi.minvenj.nl
[2]*visiting scientist*

We prefer a poster presentation

The NFI-FRITS database (Forensically Realistic Intercepted Telephone Speech) contains speech intercepted during real police investigations. This material was obtained to facilitate research on data typically encountered in forensic practice, much like the data used by Becker (2012) and Van Leeuwen and Bouten (2004) and the AHUMADA III data (Ramos et al, 2008). NFI-FRITS consists of over 4100 recordings of more than 600 speakers.

Data processing

The raw data were provided with some metadata, like the two telephone numbers involved in the telephone call, case name, etc. The audio files were split in two single channel files (a and b) and stored in a database, along with the provided metadata.
Native listeners listened to the material and removed information in the audiofiles that can identify an individual and assigned speaker names and other metadata. This was done until about five recordings were assigned to a speaker, after which the process was repeated.

Realistic data

The database consists of realistic data, meaning that the audio comes from intercepted telephone speech from police investigations. The forensic nature of the recordings and the method to label a recording with a speaker name make the truth about speaker identities in this database a truth by proxy. Nevertheless the authors feel that the method used leads to sufficiently reliable speaker identities. The data is representative of police investigations, however, the collection is not representative of casework at the NFI as this typically involves recordings where the speaker ID is disputed.

Database by numbers

The database consists of 4188 recordings and 604 speakers. There are 427 male speakers in 3120 recordings and 177 female speakers in 1068 recordings. There are 72 multilingual speakers in the database, who speak Dutch in some recordings and either Turkish, Moroccan Arabic or Berber in other recordings.

Table 1. No. recordings per language

Language	#recordings
Dutch (all varieties)	3091
Turkish	499
Moroccan Arabic	191
Berber (Tarifit)	116
Mixed	245
Other	46

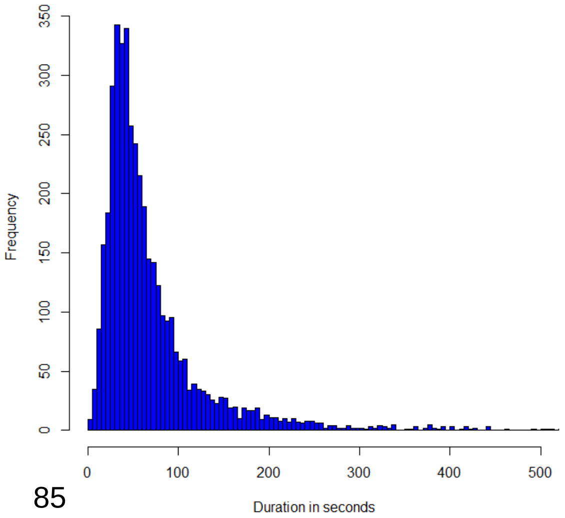


Figure 1 Histogram of measured durations of speech per recording

References

T. Becker, “Automatic forensic voice comparison (automatischer forensischer stimmenvergleich),” in The Journal of Speech Language and the Law, 2012, vol. 19, pp. 291–294.

David A. van Leeuwen and Jos S. Bouten, “Results of the 2003 NFI-TNO forensic speaker recognition evaluation,” in Proc. Odyssey 2004 Speaker and Language recognition workshop. June 2004, pp. 75–82, ISCA.

Daniel Ramos, Joaquin Gonzalez-Rodriguez, Javier Gonzalez-Dominguez, and Jose Juan Lucena-Molina, “Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish.,” in Proc. Interspeech, 2008, pp. 1493–1496.