An fMRI study on forensic phonetic speaker recognition with blind and sighted listeners

Almut Braun[1], Andreas Jansen[2], and Jens Sommer[2]

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

ABSTRACT

A forensic phonetic speaker recognition experiment with spontaneous speech samples of known and unknown speakers was carried out while listeners underwent a functional magnetic resonance imaging (fMRI) scan. In sighted participants, listening to familiar in contrast to unfamiliar speakers elicited brain activations in the right frontal pole and the left part of the cerebellum. When fMRI data of the first and the second 15 seconds of listening to familiar speakers were compared, it was found that auditory areas were significantly stronger activated in the first part and visual areas showed stronger activations in the second part. In two blind participants, there were no brain activations which were stronger in the first compared to the second 15 seconds of listening to familiar speakers' voice samples. When the second part was compared to the first, also blind listeners showed stronger activations in visual areas.

Keywords: forensic phonetics, speaker recognition, fMRI, neuroimaging, blind listeners

1. INTRODUCTION

The present fMRI study was carried out in order to investigate the neurobiological underpinnings of listeners' performance in a forensic phonetic speaker recognition test.

Over the past years, a couple of psychological neuroimaging studies have been carried out on the subject of voice recognition; however, these studies focused on basic research in the realm of human voice processing: other methods and stimuli were used than those one would typically select for a study designed from a forensic phonetic point of view. For example, these earlier studies used simple voice discrimination tasks [25, 2], mixed stimuli of male, female and sometimes also children's voices [16, 28, 13] or unnatural speech samples such as resynthesized speech [20] and simple vowels/syllables [2, 20].

Since the present study is assumed to be the first in which a complex forensic phonetic speaker recognition experiment is carried out in combination with fMRI, preliminary tests were necessary beforehand in order to ascertain the overall feasibility of the project. Details on the latter are provided in the methods section.
In short, the study addresses the following research questions: 1) Can a complex speaker recognition experiment be carried out within an MR scanner? 2) Which brain regions are active when listeners perform a complex speaker recognition experiment with either familiar speakers or speakers who have been heard only once before? 3) Do listeners' brain activations differ for recognized and not recognized speakers? 4) Why are some listeners better at recognizing people by their voices than others? 5) Do activation patterns differ between particular listener groups (e.g. blind and sighted listeners)?

2. PRIOR RESEARCH

Since voice-selective areas have been detected bilaterally along the upper bank of the superior temporal sulcus (STS) in the human auditory cortex [5,1] and in further brain regions [25, 3, 21], a growing number of neuroimaging studies have been published on voice recognition. It was found that familiar and unfamiliar voices are processed in at least partially dissociated functional pathways [4] and that processes of visual-auditory interaction take place in tasks of human person recognition [29]. Nevertheless, the nature of representations for complex speaker recognition is still unknown [27, 22]. Over the past few years, neuroimaging studies on voice recognition have also been carried out with blind listeners after several behavioral tests had shown enhanced voice recognition abilities in blind listeners [7, 8, 26] – some studies, though, did not find blind listeners to be superior to sighted controls [9, 30].

In an fMRI study in which data from vocal and non-vocal stimuli were compared to baseline, blind listeners showed activations in occipital brain regions (which are usually associated with visual processing) and relative to the group of sighted

listeners reduced activation in auditory areas, whereas sighted listeners showed the opposite pattern [13]. When hemodynamic (blood oxygenation level-dependent, BOLD) responses to vocal and non-vocal stimuli were compared to each other, it was found that congenitally blind listeners showed significantly stronger activations in the left STS to vocal than non-vocal stimuli – a pattern that was absent in late blind as well as sighted listeners. Furthermore, STS activations found in the pooled group of congenitally and late blind listeners correlated with those listeners' results in an offline performed speaker recognition test [13]. Hölig et al. presented sighted, congenitally blind [14] and late blind [15] listeners with an fMRI speaker discrimination task and found that for person-incongruent trials, congenitally blind and late blind listeners had significantly stronger activations in the right anterior fusiform gyrus – an area which is also associated with face processing [23]. This was not the case for sighted listeners.

In a pre-scanning voice learning phase, blind listeners learned the voices significantly faster than sighted controls [14,15] and congenitally blind listeners also obtained significantly better results in a pre-scanning speaker recognition test [14]. In an EEG study using the same priming paradigm as the two aforementioned studies, congenitally blind – but not sighted – listeners showed a significantly enhanced negativity 100-160 ms after the onset of the second stimulus in person-incongruent compared to person-congruent trials [10].

3. A NEW APPROACH

Combining forensic phonetic speaker recognition experiments with neuroimaging techniques such as fMRI offers new perspectives, but also raises some important issues that need to be addressed. For instance, without appropriate shielding, the noisy environment of an MR scanner makes it completely impossible to meet – or at least approach – the requirements of a calm place to carry out a proper forensic speaker recognition experiment [19]. Secondly, the exact point in time where a listener recognizes a particular speaker is not well defined and it will also vary between listeners and vary between speakers. Active reporting of a successful recognition, for example by pressing a button, would shift the attention of the subjects and potentially contaminate the results [11]. As the use of only one target speaker whose voice was only presented once does not produce enough fMRI data, a single-presentation voice lineup cannot be converted one-to-one into an fMRI study.

Voice recognition abilities of blind listeners have already been investigated from a neuropsychological point of view in a couple of fMRI studies [13, 14, 15]; however, the same methodological problems as mentioned before exist when the results of those studies are tried to be interpreted in a forensic phonetic context. The present study is an attempt to conduct a complex speaker recognition experiment (with blind and sighted listeners) within an fMRI setting. Interdisciplinary research in this area is crucial for answering the questions posed in this paper.

4. METHOD

The experiment consisted of two parts, separated by 5 minutes of anatomical scans, and a block design was used for stimulus presentation. In part A, a sound file with 15 spontaneous voice samples of different male native speakers of German (aged 45-63) was played to the listeners lying inside a 3-Tesla high-field MR scanner. All speech samples were good-quality recordings taken from German talk-shows: 10 of the samples came from famous speakers which were supposed to be recognized easily and 5 samples were from unknown speakers. Voice samples (duration: 30 seconds, respectively) were randomized and followed by a silent interval of 10 seconds.

After completion of the anatomical measurements, part B of the experiment was carried out. Here, participants listened to a second test file which consisted of 3 of the previously unknown voices from part A (the verbal content was different) as well as 6 new unknown voices in randomized order. To avoid position effects, different versions of test file A and B were used. The task while listening to the test files was in both cases to try to identify the speakers.

A pretest was carried out with 12 listeners (all female, aged 19-25, native speakers of German) who were not involved in the main experiment in order to test the familiarity of the speakers. None of the supposedly unknown speakers was recognized and all famous speakers were recognized at least once. 3 of the famous speakers were recognized by 11 out of 12 pretest listeners. The pretest should ensure that listeners are able to identify at least some of the famous speakers.

14 medical students (right-handed, 5 male, aged 20-27) and two congenitally blind listeners (right-handed, 1 male aged 18 without any residual vision, 1 female aged 36 with minimal light perception) who reported no hearing difficulties participated as listeners in the main experiment. After scanning, listeners were asked (in separate

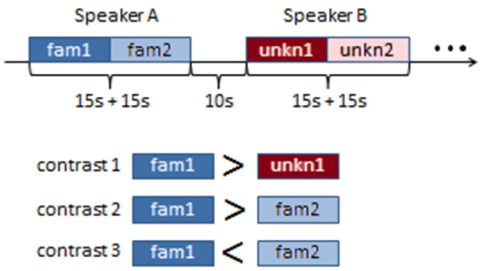
sessions) to recall as many speakers as possible from the speakers he/she had heard while undergoing the fMRI scan. Then the names of all speakers from the test files were read to the listeners in order to determine which of the speakers he/she also knew in advance. Possible answers are shown in Table 1.

Table 1: Listeners' possible answers in the post-scanning naming task. fMRI data connected with answers in bold were considered for analysis.

The participant could…
1. immediately give the speaker's name
2. give the speaker's name after questioning
3. give a description of the speaker, but no name
4. Listeners gave two names for a speaker
5. Listeners said the speaker sounded familiar
6. No recognition, but listener knows the speaker
7. The listener did not know the speaker

fMRI measurements were done with a 3T MRI scanner (Magnetom Trio, SIEMENS) with a BOLD-sensitive (blood oxygenation level-dependent) imaging sequence. Imaging parameters were: repetition time (TR) 1580 ms, echo time (TE) 30ms, flip angle (FA) 70°, 25 consecutive slices with 5mm thickness, a gap of 10%, a field of view of 192mm, and a matrix of 64x64, yielding voxels of 3x3x5mm³. The anatomical data was acquired with a MPRAGE-sequence, TR 1900ms, TE 2.26ms, inversion time 900ms, FA 9°, parallel imaging (GRAPPA) factor 2, isometric voxels of 1mm³. The analysis of fMRI data was done with FSL, FMRIB Software Library 5.0.7, Oxford [12, 17, 18]. Analyses on subject level included temporal high-pass filtering (threshold frequency 0,017 Hz) in order to remove very slow changes, smoothing with a gaussian kernel of 8mm full width half maximum (FWHM), motion correction, slice timing correction, registration of the functional data to individual anatomical data and to the ICBM152 brain (MNI 152, Montreal Neurological Institute). Group analysis was performed with a mixed effects model (FLAME, FMRIB's local analysis of mixed effects) and resulting statistic images were thresholded using clusters determined by $z > 2.3$ and a cluster significance threshold of $p = 0.05$, (family wise error correction, FWE). For the analysis, listener's fMRI data for listening to familiar and unknown speakers was split into halves of 15 seconds in order to make further comparisons (see Figure 1).

Figure 1: Test file A (voice samples from familiar/ unknown speakers) and contrasts used for analysis.



5. RESULTS

Figure 2 (left column) shows the contrast between the first 15 seconds of listening to familiar speakers and the first 15 seconds of listening to unknown speakers. It shows that the right frontal pole (top) and the cerebellum (bottom) are more activated while listening to familiar speakers.

Figure 2 (middle and right column) displays contrasts for the split half analysis, i.e. contrasts 2 and 3. Significantly stronger activations were found during the first 15 seconds bilaterally in the superior temporal sulcus (STS), the superior temporal gyrus (STG), Heschl's gyrus, the central opercular gyrus and the temporal poles (Figure 2, middle column). During the second 15 seconds, the frontal poles, cingulate gyrus, precuneus, lateral occipital cortex, right angular gyrus, putamen, right superior frontal gyrus and the left temporal pole showed higher activation (Figure 2, right column).

Figure 2: Activations of contrasts 1 (left column), 2 (middle column) and 3 (right column) in sighted listeners. Numbers refer to MNI coordinates. A=anterior, S=superior, R=right

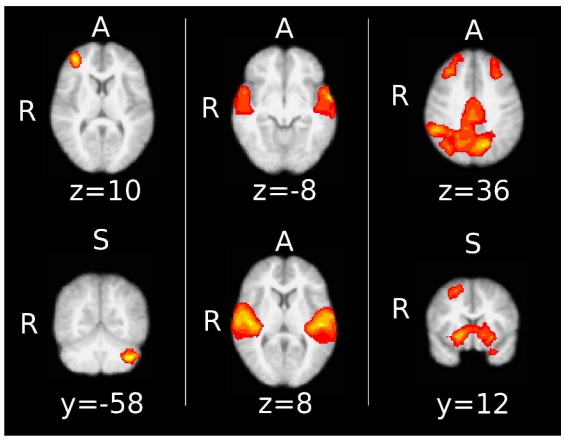
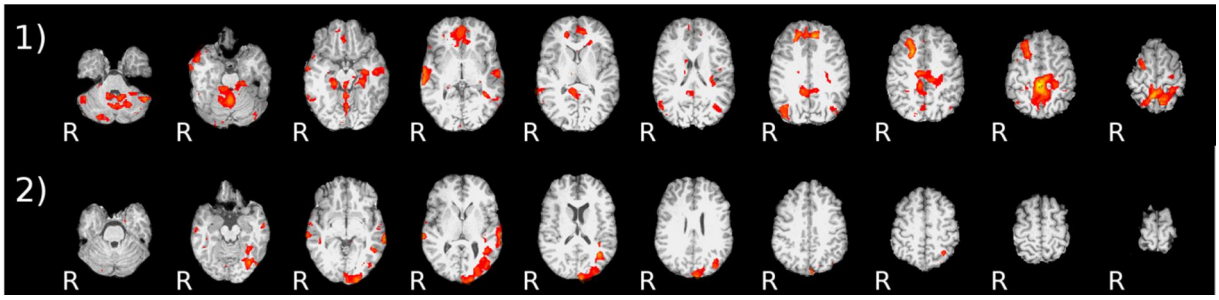


Figure 3: Activations of contrast 3 in two congenitally blind listeners: 1) male blind, 2) female blind participant.



6. DISCUSSION

When fMRI data of sighted participants listening to known vs. unknown voices were compared, stronger brain activations were found in the in the right frontal pole and the cerebellum. The former area has been associated with episodic memory functions [25] and connecting vocal attributes to visual speaker representations [16], the latter is said to play a role in the modulation of a variety of linguistic functions, e.g. word retrieval and metalinguistic abilities [24].

For sighted participants, comparing the neuroimaging data of the first 15 seconds to the second 15 seconds of listening to familiar speakers showed significantly stronger bilateral activations in the superior temporal sulcus (STS), the superior temporal gyrus (STG), Heschl's gyrus, the central opercular gyrus and the temporal poles (see Figure 2, middle column). These areas are associated with auditory analysis and cognitive control processes [6]. For blind participants, the aforementioned contrast did not reveal any significant differences between the first compared to the second 15 seconds.

When the comparison was reversed, i.e. when the second 15 seconds of sighted participants listening to familiar speakers were compared to the first 15 seconds, the following brain areas showed a stronger activation: the frontal poles, the cingulate gyrus, the precuneus, the lateral occipital cortex, the right angular gyrus, the putamen, the right superior frontal gyrus and the left temporal pole (see Figure 2, right column). Functions associated with these areas are monitoring decisions, episodic memory, coordinating sensory input with emotions, visuospatial imagery, object recognition, visuospatial processing, reinforcement learning and cognitive control processes [6]. For the two congenitally blind participants, the latter contrast revealed large differences in brain activation. More blind participants are necessary in order to run a group analysis.

Stronger brain activations found for listening to complex natural speech samples of familiar speakers compared to unknown speakers were found in the right frontal pole in this study; however, a PET study found activations in the left frontal pole [25]. For familiar speakers, auditory processing was stronger within the first 15 seconds compared to the second 15 sec. of a given stimulus. Regarding the latter result, it is possible that sighted listeners visualized the familiar speakers after successful voice processing and recognition. Interestingly, the occipital cortex was also activated in congenitally blind participants. Note that the present study is a feasibility study in which only two congenitally blind participants are included so far. More blind participants are needed to carry out group analyses.

Furthermore, it would be interesting to see whether the chosen paradigm can also be used for testing whether there is a difference between speakers heard once before and completely unknown speakers (test file B).

7. FUTURE PROSPECTS

In order to fully understand the complex processes of human speaker recognition, far more studies are needed. From a forensic point of view, it would be interesting to investigate the neural representations of once heard voices in more detail to analyze the question why some voices are easier to remember than others and e.g. whether unfamiliar but highly characteristic voices are similarly processed as familiar voices. Results of future studies might also be useful to create a voice recognition training program with neurofeedback for people who want to specialize in forensic phonetics. Furthermore, it would be interesting to compare fMRI data of forensic phonetic experts and lay listeners since experts performed significantly better than blind and sighted lay listeners in a behavioral speaker recognition task [9].

Another question worth exploring would be whether it is possible to identify determinants of voice misidentification [27]. This could, for instance, help to exonerate innocent suspects in the future. Furthermore, it would be interesting to examine whether "nonconscious speaker recognition" is possible. In other words, to see whether the brain responds to a particular voice as if it was recognized, but without the listener being consciously aware of this fact. Finally, neurological correlates of speaker recognition might someday also help to predict the reliability of an earwitness. Although the ideas mentioned above sound interesting and promising, the authors want to stress the point, that any diagnostic

application of fMRI used for forensic purposes needs to be thoroughly tested and critically assessed beforehand. Further research in this area may help to avoid and prevent the overhastily and uncritically adoption of "brain prints" and the misuse of fMRI techniques in the field of forensic phonetics.

8. REFERENCES

[1] Andics, A., Gál, V., Vicsi, K., Rudas, G., Vidnyánszky, Z. 2013. FMRI repetition suppression for voices is modulated by stimulus expectations. *Neuroimage* 69, 277–283.

[2] Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z. 2010. Neural mechanisms for voice recognition. *Neuroimage* 52, 1528–1540.

[3] Arnott, S.R., Heywood, C.A., Kentridge, R.W., Goodale, M.A. 2008. Voice recognition and the posterior cingulate: An fMRI study of prosopagnosia. *Journal of Neuropsychology* 2, 269–286.

[4] Belin, P., Fecteau, S., Bédard, C. 2004. Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences* 8, 129–135.

[5] Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B. 2000. Voice-selective areas in human auditory cortex. *Nature* 403, 309–312.

[6] Bösel, R. M. 2006. *Das Gehirn. Ein Lehrbuch der funktionellen Anatomie für die Psychologie*. Stuttgart:W. Kohlhammer GmbH.

[7] Braun, A. 2012. Speaker Recognition Ability of Blind and Sighted Subjects. *The International Journal of Speech, Language and the Law* 19(2), 159–187.

[8] Bull, R., Rathborn, H., Clifford, B.R. 1983. The Voice-Recognition Accuracy of Blind Listeners. *Perception* 12, 223–226.

[9] Elaad, E., Segev, S., Tobin, Y. 1998. Long-Term Working Memory in Voice Identification. *Psychology, Crime and Law* 4, 73–88.

[10] Föcker, J., Best, A., Hölig, C., Röder, B. 2012. The superiority in voice processing of the blind arises from neural plasticity at sensory processing stages. *Neuropsychologia* 50, 2056–2067.

[11] Frässle, S., Sommer, J., Jansen, A., Naber, M., Einhäuser, W. 2014. Binocular rivalry: frontal activity relates to introspection and action but not to perception. *Journal of Neuroscience* 34(5), 1738–1747.

[12] FSL, FMRIB Software Library 5.0.7, Oxford http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

[13] 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, 2967–2974.

[14] Hölig, C., Föcker, J., Best, A., Röder, B., Büchel, C. 2014. Brain systems mediating voice identity processing in blind humans. *Human Brain Mapping* 35(9), 4607–4619.

[15] Hölig, C., Föcker, J., Best, A., Röder, B., Büchel, C. 2014. Crossmodal plasticity in the fusiform gyrus of late blind individuals during voice recognition. *Neuroimage* 103, 374–382.

[16] Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Suqiura, M., Fukada, H., Itoh, K., Kato, T., Nakamura, A., Hatano, K., Kojima, S., Nakamura, K. 1997. Vocal identification of speaker and emotion activates different brain regions. *Neuroreport* 8, 2809–2812.

[17] Jenkinson, M., Smith, S.M. 2001. A Global Optimisation Method for Robust Affine Registration of Brain Images. *Medical Image Analysis* 5(2), 143–156.

[18] Jenkinson, M., Bannister, P., Brady, M., Smith, S. 2002. Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *Neuroimage* 17(2), 825–841.

[19] Kersthold, J.H., Jansen, E.J.M., van Amelsvoort, A.G., Broeders, A.P.A. 2005. Richtlijnen auditieve confrontatie. *Politiekunde* 8, 3–31.

[20] Kreitewolf, J., Gaudrain, E., von Kriegstein, K. 2014. A neural mechanism for recognizing speech spoken by different speakers. *Neuroimage* 91, 375–385.

[21] Latinus, M., Crabbe, F., Belin, P. 2011. Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex* 21, 2820–2828.

[22] Mathias, S.R., von Kriegstein, K. 2014. How do we recognise who is speaking? *Frontiers in Bioscience (scholar edition)* 1(6), 92–109.

[23] Morris, J.P., Pelphrey, K.A., McCarty, G. 2005. Cognitive Neuroscience Society. Poster. URL: http://www.biac.duke.edu/library/posters/2005_morris_cns.pdf

[24] Murdoch, B.E. 2009. The cerebellum and language: Historical perspective and review. *Cortex* 46, 858–868.

[25] Nakamura, K. Kawashima, R., Suqiura, M., Kato, T., Nakamura, A., Hatano, K., Naqumo, S., Kubota, K., Fukuda, H., Itoh, K., Kojima, S. 2001. Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia* 39(10), 1047–1054.

[26] Röder, B., Neville, H. 2003. Developmental functional plasticity. In: Grafman, J., Robertson, I.H. (eds.), *Handbook of Neuropsychology. IX. Plasticity and Rehabilitation.* Oxford: Elsevier Science, 231–270.

[27] Schweinberger, S.R., Kawahara, H., Simpson, A.P., Skuk, V.G., Zäske, R. 2014. Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 5(1), 15–25.

[28] von Kriegstein, K., Giraud, A.L. 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* 22, 948–955.

[29] von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L. 2005. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience* 17(3), 367–376.

[30] Winograd, E., Spence, M.J., Kerr, N.H. 1984. Voice recognition: effects of orienting task, and a test of blind versus sighted listeners. *American Journal of Psychology* 97, 57–70.