Is it still speech? Different processing strategies in learning to discriminate stimuli in the transition from speech to non-speech including feedback evaluation

Tina Weis\textsuperscript{a}, Christoph M. Krick\textsuperscript{b}, Wolfgang Reith\textsuperscript{b}, Thomas Lachmann\textsuperscript{a,c,}\textsuperscript{*}

\textsuperscript{a} Cognitive and Developmental Psychology Unit, Center for Cognitive Science, University of Kaiserslautern, Kaiserslautern, Germany
\textsuperscript{b} Clinic of Diagnostic and Interventional Neuroradiology, Saarland University Hospital, Homburg, Germany
\textsuperscript{c} University of Leuven, Leuven, Belgium

\section*{ARTICLE INFO}

Keywords:
Morphed stimuli
Discrimination task
Feedback anticipation
Auditory processing
fMRI
Spectral rotated speech

\section*{ABSTRACT}

Processing of speech was investigated by using stimuli gradually changing from speech (vowels) to non-speech (spectral rotated vowels). Stimuli were presented in descending levels of vocalization blends, from pure speech to non-speech, through step-wise combinations, resulting in ambiguous versions of the sounds. Participants performed a two-alternative forced choice task: categorization of sounds were made according to whether they contained more speech or non-speech. Performance feedback was presented visually on each trial. Reaction times (RT) after sound presentation, and functional magnetic resonance imaging (fMRI) data during auditory and visual processing, were analyzed. RT data suggested individual differences with a distinct group, \textit{good performers}, functioning better in distinguishing stimuli with a higher degree of ambiguous blends compared to \textit{poor performers}, who were not able to distinguish these stimuli correctly. fMRI data confirmed this finding. During auditory stimulation, \textit{good performers} showed neural activation in the ventral auditory pathway, including the primary auditory cortex and the anterior superior temporal sulcus (responsible for speech processing). \textit{Poor performers}, in contrast, showed neural activation in the dorsal auditory pathway, including the bilateral superior temporal gyrus. Group differences were also found for visual feedback processing. Differences observed between the groups were interpreted as reflecting different neural processing strategies.

\section*{1. Introduction}

\subsection*{1.1. Processing of speech and non-speech}

What makes an "/a/" an "a"? The speech signal is a very complex acoustic stimulus, with many possible variations, but it is still recognizable as “speech”. Phonemes, syllables and words can be spoken by males and females, in different emotional conditions, in different languages (or regional dialects) and under different environmental conditions, such as noise, and still be understood by a listener (e.g., Holt & Lotto, 2010; Liberman, 1996; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1989). Yet, how are we still able to differentiate speech from non-speech sounds?

There is a large number of studies focusing on the perception and processing of speech versus non-speech (e.g., Blesser, 1972; Christmann, Berti, Steinbrink, & Lachmann, 2014; Firszt, Ulmer, & Gaggl, 2006; Lazard, Collette, & Perrot, 2012; Niätäinen & Winkler, 1999). Most of these used either a categorization task or a speech-in-

 noise paradigm.

In a categorization task, speech sounds have to be distinguished from non-speech sounds (Christmann et al., 2014; Kuuluvainen et al., 2014; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Olesser, Zimmerman, van Meter, & Rauschecker, 2007). Importantly, to investigate differences in processing speech versus non-speech and the underlying neuronal mechanisms, stimulus complexity between the categories should be comparable. However, in most of these studies the stimulus complexity was not balanced between categories. With respect to spectro-temporal characteristics, the non-speech stimuli used, such as sinusoidal tones or simple tone patterns (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Démontet et al., 1992; Jäncke, Wüstenberg, Scheich, & Heinze, 2002; Lachmann, Berti, Kujala, & Schröger, 2005; Poeppel et al., 2004; Vououmanos, Kiehl, Werker, & Liddel, 2001), noise bursts (Zatorre, Evans, Meyer, & Gjedde, 1992), or environmental sounds (Giraud & Price, 2001), were less complex than the used speech stimuli.

Spectrally rotated noise-vocoded speech (Narain et al., 2003; Scott, ...
Blank, Rosen, & Wise, 2000) represents an appropriate way to create non-speech control sounds of equal complexity compared to speech sounds. However, due to higher order linguistic analyses, including lexical, semantic, and syntactic processing, spectrally and/or temporally degraded speech sounds may still be understandable (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Therefore, in the present categorization study, isolated vowels as speech sounds, and their spectrally rotated counterparts as non-speech sounds (spectrally rotated vowels; Christmann et al., 2014) were used as stimuli. This results in speech and non-speech sounds of comparable stimulus complexity. At the same time, there are no higher order linguistic influences, as evident from previous behavioral (Vandermosten et al., 2010, 2011) and neuroimaging studies (Liebenthal et al., 2005; Obleser et al., 2006; Scott, Rosen, Lang, & Wise, 2006).

The second paradigm, often applied to investigate speech perception and processing, uses speech stimuli that are embedded into a variety of background noises (“speech-in-noise”, e.g., Klatte, Lachmann, & Meis, 2010). The limitation here, however, is that noise signals are artificial (e.g., white noise) or mixed speech (Bronkhorst, 2015; Cherry, 1953).

In the present study, both paradigms, categorization and speech-in-noise, were combined, using naturally spoken vowels and their spectrally rotated counterparts. Consequently, speech and non-speech stimuli have the same complexity in terms of their spectro-temporal characteristics. Moreover, gradually changing blended speech and non-speech stimuli were used in the present study. Each individual sound contained either more speech or more non-speech. This allowed for better comparability between the stimulus categories, but made the task harder to solve for the participants. Blended stimuli were used in previous studies (Oses, Hugdahl, Hjelmervik, & Specht, 2011; Oses, Hugdahl, & Specht, 2011; Specht, Oses, & Hugdahl, 2009). In these studies, in contrast, the speech stimuli that morphed with white noise were less complex than their speech stimuli (see Fig. 1 in Specht et al., 2009).

In the present functional magnetic resonance imaging (fMRI) study, the processing of blended speech and non-speech stimuli in a two-alternative forced choice task was investigated. Both the speech and the non-speech category contained either pure sounds, i.e. pure speech or pure non-speech that were easy to discriminate (categorical), or blended sounds that differed gradually in proportion of speech and non-speech components, which were more difficult to discriminate (speech-in-noise).

The study aimed to determine whether there is either an internal representation of gradually changing stimuli from speech to non-speech, or whether there are two distinct representational categories. Participants had to judge whether the presented stimulus sounded more like speech or more like non-speech. Since categorization of the morphed stimuli tends to be difficult, trial-based performance feedback was included. Depending on the correctness of their categorization, participants received either positive or negative feedback that was presented visually. In some trials, irrespective of the response, a neutral visual feedback was given. A pilot study without performance feedback showed that most participants were able to categorize morphed stimuli correctly, while other participants had greater difficulties. Since it was not clear whether these participants simply did not understand the instructions correctly or whether they were, in fact, unable to differentiate the different amounts of speech in the sounds, performance feedback was included. Participants were expected to learn the categorization based on feedback evaluation.

Regarding the neural mechanisms underlying speech processing, it is commonly assumed that processing is organized hierarchically, in a dorsal and a ventral pathway (Rauschecker & Scott, 2009). In the dorsal pathway, primary auditory areas in the bilateral superior temporal lobe are most commonly activated following acoustic stimuli (Belin et al., 2000; Binder et al., 2000; Mummery, Ashburner, Scott, & Wise, 1999; Poeppel et al., 2004; Wise et al., 1991; Zatorre et al., 1992), whereas, in the ventral pathway, the left superior temporal gyrus and left superior temporal sulcus are sensitive to the spectro-temporal complexity and intelligibility of acoustic sounds (Binder et al., 2000; Giraud et al., 2004; Mummery et al., 1999; Poeppel, 2003; Rauschecker & Tian, 2000; Tian, Reser, Durham, Kustov, & Rauschecker, 2001). Direct comparison between stimuli with similar complexity, but different intelligibility, resulted in stronger activation in the left anterior and middle superior temporal cortex for speech compared to spectrally rotated non-speech sounds (Liebenthal et al., 2005; Narain et al., 2003; Scott et al., 2000; Scott et al., 2006). However, there were no differences during processing of the different stimulus categories in the bilateral dorsal superior temporal gyrus, close to the Heschl’s gyrus (Liebenthal et al., 2005).

1.2. Feedback anticipation

The current task is also a discrimination learning task as the participants were presented with feedback on their performance. Thus, feedback anticipation during discrimination of blended speech and non-speech stimuli was investigated. Numerous studies have convincingly shown that learning the association between a stimulus and possible reward or punishment allows us to anticipate the respective feedback (Bakin, South, & Weinberger, 1996; Beitel, Schreiner, Cheung, Wang, & Merzenich, 2003; Blake, Heiser, Caywood, & Merzenich, 2006; Condon & Weinberger, 1991; Edeline & Weinberger, 1991; Ohl, Scheich, & Freeman, 2001).

On one hand, feedback can be aversive, such as fear induced by a minor electric foot shock (Thiel, Bentley, & Dolan, 2002; Thiel, Friston, & Dolan, 2002). On the other hand, feedback can be appetitive, such as happiness induced by receiving a monetary reward (Puschnig, Brechmann, & Thiel, 2013; Weis, Puschnig, Brechmann, & Thiel, 2012). In both cases, the association between an auditory stimulus and the respective feedback led to enhanced activation in the auditory cortex, modulated by either the cholinergic (aversive) or the dopaminergic (appetitive) system (see also Baö, Chan, & Merzenich, 2001; Guitart-Masip et al., 2012; Kisley & Gerstein, 2001; Knutson, Fong, Adams, Varner, & Hommer, 2001; Wittmann et al., 2005).

In the current study, participants had to learn the correct categorization of blended stimuli into either a “more speech” or a “more non-speech” category. They received positive feedback for correct responses and negative feedback for incorrect responses.

The consequence of using a blended stimulus is that some are more ambiguous than others. Those that are least ambiguous are easier for participants to categorize and thus they may anticipate positive feedback. However, for more ambiguous stimuli, categorization proves difficult and the participants may anticipate negative feedback. Thus, feedback anticipation in the current study is more a self-monitoring of performance that helps with solving the task correctly than an instrumental conditioning, because correct answers lead to a positive response and not a stimulus category as such. Hence, even if initially, the participants incorrectly categorize most of the ambiguous stimuli as “non-speech”, feedback will enable them to learn the correct categorization.

1.3. Feedback delivery

While the effects of feedback anticipation in sensory cortices have been investigated in many studies, less is known about the effects of feedback processing in sensory cortices when different modalities (i.e., auditory and visual) are used. Some studies investigating the effect of reward delivery have linked neural activations in the sensory cortices to previous stimulus presentation, but not to the feedback as such (Pleger, Blankenburg, Ruff, Driver, & Dolan, 2008; Weil et al., 2010; Weis et al., 2012; Weis, Brechmann, Puschnig, & Thiel, 2013; Weis, Puschnig, Brechmann, & Thiel, 2013). Weis, Brechmann, et al. (2013) found...
enhanced neural activation in the auditory cortex during visual feedback presentation using an instrumental learning task in which participants had to determine the reward-predicting feature in frequency-modulated auditory stimuli. Interestingly, this activation was dependent on participants’ expectations. Enhanced activation in the auditory cortex only occurred in participants who learned the tone-reward association. Hence, expectancy plays a major role during feedback processing, and leads to a kind of “teaching signal” in sensory cortices.

In a follow-up study (Weis, Puschmann, et al., 2013) that used an auditory duration discrimination task, two separate sessions were used for reward and punishment. Confirming the previous results, enhanced neural activity for gaining a reward or avoiding losing money was found in the auditory cortex. In summary, activation in sensory cortices occurred as a consequence of receiving a reward or avoiding punishment. As shown there, participants were able to establish a discrimination threshold. Therefore, in the current study, the discrimination of blended speech and non-speech stimuli was investigated and feedback for correct and incorrect responses as well as neutral feedback (independent of response), randomly distributed over trials, implemented in one session was provided to the participants.

1.4. Current study

Two different processes were investigated: (1) auditory stimulus discrimination, i.e., differences in processing of blended speech and non-speech stimuli, including feedback anticipation; and (2) visual feedback processing, i.e., different activation patterns for positive, negative, and neutral feedback.

In line with the previous behavioral results (Weis, Puschmann, et al., 2013), it was anticipated that differences would be found in reaction times (RT) between correct and incorrect responses, because participants must learn which stimuli should be categorized as containing “more speech” or “more non-speech”. Due to the performance feedback, an improvement in performance from the first half (learning) compared to the second half (solidification) of the trials was expected, even if some of the stimuli are harder to discriminate than others.

During auditory stimulus discrimination, it was anticipated that differences between stimuli that contained more speech and stimuli that contained more non-speech in the left anterior and middle superior temporal cortex would be found, as described in earlier fMRI studies (Liebenthal et al., 2005; Narain et al., 2003; Scott et al., 2000; Scott et al., 2006). However, it was also anticipated that the resulting data might result in less stable effects, since blended speech and non-speech stimuli were used, which cannot be clearly categorized.

During feedback presentation, it was anticipated that there would be stronger activation of the auditory cortex for correct compared to neutral or incorrect feedback conditions. This would permit further investigation of the underlying processes of the reactivation of sensory cortices during feedback delivery.

2. Methods

2.1. Participants

A total of 30 native German speaking students of the University of Kaiserslautern (17 males, age range: 19–30 years, average age: 23.8 ± 2.6 years) participated in the study. They were all right-handed as indexed by a handedness inventory (Oldfield, 1971), reported normal hearing and vision, and were either paid for their participation or received course credits. The study was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013), and the experiments were approved by the local ethics committee. Written informed consent was obtained from all participants who were further screened for their MRI compatibility before entering the scanner room. None of the participants had to be excluded due to head movements during fMRI measurement (head movement measured as absolute displacement across the run < 3 mm and as scan-to-scan displacement < 1 mm). Two participants had to be excluded due to technical problems during data acquisition. During analysis of the behavioral data, five participants, whose answers were too slow in more than one-third of the trials, were excluded. Thus, 23 participants remained for further analyses (12 males, average age: 23.8 ± 2.8 years).

2.2. Auditory stimuli

In German, there are fourteen vowel monophthongs that can be grouped into seven pairs, the members of which differ exclusively with respect to tenseness (tense vs. lax; Wiese, 2000), i.e. vowel length (long vs. short; Lühr, 2000). In the current study, auditory stimuli consisted of a selected subset of the stimuli created by Christmann et al. (2014): the German vowels /a/ and /a:/ and their spectrally rotated counterparts.

The two naturally spoken vowels /a/ and /a:/ were pronounced in isolation by a female native speaker of German. The recording was performed in a sound-attenuated cabin by using Adobe Audition 1.5 with a sampling rate of 44 kHz. The stimuli were adjusted in their respective duration (75 ms for /a/ and 145 ms for /a:/) according to the method defined by Groth, Lachmann, Riecker, Muthmann, and Steinbrink (2011). Consequently, only the steady-state part of the vowels was used to obtain static spectral information. Vowel length was identified by visual inspection of spectrograms (formant movements) and waveforms using the phonetics program “Praat” (Boersma, 2002; Boersma & Weenink, 2013). The Pitch Synchronous Overlap and Add (PSOLA) algorithm was used to adjust vowel length without distorting the spectral properties. All stimuli included an amplitude modulated fade-in and a fade-out time of 5 ms. A pre-emphasis high-pass filter was applied for compensation of inconsistencies due to differential sensitivity of the human ear to sound energies at different frequencies (Killion, 1978; Scott et al., 2000). This procedure improves the matching of the power spectra between the speech sound and its spectrally rotated counterpart (Obleser, Scott, & Eulitz, 2006).

The general procedure to create spectrally rotated sounds was first described by Blesser (1972), Christmann et al. (2014) produced the stimuli with a MATLAB script for stimulus processing (see http://www.phon.ucl.ac.uk/resource/software.php), provided by Scott et al. (2000). In contrast to the procedure described by Blesser (1972), Christmann et al. (2014) improved the spectrally rotated stimuli for two reasons. First, previous studies used a cut-off frequency for low-pass filtering of 4000 Hz (Davids et al., 2011; Evans et al., 2014; Scott et al., 2000, 2006), because the usable voice frequency band ranges from 500 and 4000 Hz (Wilmanns & Schmitt, 2002). This is particularly true for vowel stimuli, as used in the current experiment. Note that for distinguishing fricatives, higher frequency bands might also be important. However, while this kind of low-pass filtering ensures the intelligibility of the speech sounds (Scott & Wise, 2004), the naturalness could be severely reduced (Moore & Tan, 2003). Therefore, by using the unfiltered spectrum, Christmann et al. (2014) ensured that the speech signal sounded natural. Second, since the study by Blesser (1972) showed that the spectral rotation of fricatives, nasals, and plosives resulted in other speech-like impressions (because in those instances, the higher frequency bands are important as well), Christmann et al. (2014) only used vowels to ensure a non-speech-like impression. Finally, the root mean squared amplitudes of the sounds were matched to ensure equal loudness.

As described above, frequencies above 4000 Hz were unaltered in both types of stimuli (see Fig. 1, black line) to allow for a more complete spectrum (unfiltered) and more naturalness of the stimuli. When spectrally rotated, a predefined rotation frequency (2000 Hz, see Fig. 1, dotted line) was established and all frequencies that fell below it, were turned into high frequencies and vice versa (see Fig. 1, non-speech). Thus, the overall spectro-temporal pattern (including complexity) of the rotated stimuli remained the same as in the original speech sound, but intelligibility was impaired.
Each vowel (/a/ and /a:/) and its spectrally rotated counterpart were blended into each other with the goal of having a range of stimuli with different levels of blending between speech and non-speech. In other words, stimuli that are easier to discriminate for the participants (pure speech: vowel; pure non-speech: spectrally rotated vowel) as well as stimuli that are harder to discriminate (blended stimuli) were developed. Therefore, as a first step in the blending procedure, the respective frequencies in the signals were identified through a fast Fourier transformation. Secondly, the amplitudes in each frequency were multiplied with the corresponding specific ratio to have a different amount of speech and non-speech within each stimulus, i.e. if a resulting stimulus contained 10% speech and 90% non-speech, the amplitude of the speech sound was multiplied with 0.1 and the amplitude of the non-speech sound with 0.9. Finally, the resulting “10–90” stimuli were built up by using the inverse fast Fourier transformation. Using this procedure, 21 different stimuli were produced in 5% steps running from pure speech sounds (100% speech and 0% non-speech, see Fig. 1, left: pure speech) to pure non-speech sounds (0% speech and 100% non-speech, see Fig. 1, middle: pure non-speech), for the short /a/ and long /a:/ stimuli, respectively. Due to issues with experimental timing, only the following were used out of the 21 stimuli (ratio of speech – non-speech): 100–0, 85–15, 70–30, 60–40, 55–45, 50–50, 45–55, 40–60, 30–70, 15–85, and 0–100 (see Fig. 1, right: 50–50 stimulus). Stimuli that contained either more speech or more non-speech were easier for the participants to differentiate and were presented in a larger step-size (i.e., 100%, 85%, 70%). Those that were harder to differentiate for the participants were presented with smaller steps (60%, 55%, 50%, 45%, 40%) to ensure that the task is difficult enough to obtain a representative number of error trials. The sound level of the stimuli was individually adjusted under scanner noise.

2.3. Visual stimuli

In addition to speech-non-speech discrimination, feedback processing was investigated including positive, negative, and neutral feedback. Previous studies showed that there is no difference in neural activity related to the amount of money paid as reward (Weis, Puschmann, et al., 2013) and that feedback processing was independent of the dopamine level (Weis, Brechmann, et al., 2013). Therefore only informative feedback instead of a monetary reward was used. Thus, the following emojis with different facial expressions and different colors were used: a green happy emoji, a red sad emoji and a yellow neutral emoji.

2.4. Task

At the beginning of each trial, in the auditory two-alternative forced choice task, an auditory stimulus was presented. Participants had to indicate, by button presses with their right index and middle fingers, if the stimulus in their perception sounded more like speech or more like non-speech. The respective button assignment was counterbalanced across participants. Note that the boundary for the judgment of speech and non-speech, and hence the decision about correct and incorrect responses, was the 50–50 stimulus. A pilot study in which participants had to judge the stimuli into “more speech” and “more non-speech” without feedback (and therefore no decision criterion), revealed an individual threshold for speech and non-speech stimuli to be between the 40–60 and the 60–40 stimulus for most of the participants. Some of them were not able to differentiate the stimuli and judged most of the them as “non-speech”, even if there was still a considerable amount of speech included. Including the performance feedback, participants were able to learn the categorization and prevent misunderstanding of task instructions. After a temporal jitter ranging from 1.78 s to 7.12 s in steps of 1.78 s, visual feedback was presented in the middle of the screen, according to participants’ performance. For correct answers, they received a green happy emoji and for incorrect answers a red sad emoji. In 25% of the trials, which were randomly distributed throughout the experimental procedure, and in all trials in which the 50–50 stimulus was presented, a yellow neutral feedback emoji was presented irrespective of correctness. Participants were forced to answer as quickly and correctly as possible. For RTs longer than one second, they received a red emoji, independent of their performance. This time-out was induced to ensure fast answers and to allow a representative amount of error trials. The feedback remained on the screen for 500 ms and was then replaced by the fixation cross which was presented on the screen during all delays and during the presentation of the auditory stimuli. The inter-trial interval ranged from 3.6 s to 8.9 s in steps of 1.78 s. The long delays between stimulus and feedback presentation, as well as the inter-trial interval, enables separation of neural responses to auditory speech-non-speech discrimination and visual feedback processing. Each of the eleven stimuli for each of the two stimulus lengths was repeated eight times and randomly presented throughout the experiment, resulting in 176 trials (see Fig. 2). All experimental control software was programmed in MATLAB (The MathWorks, Inc., Natick MA, USA), using Cogent 2000.

2.5. FMRI data acquisition

The fMRI data acquisition was performed using a 3T Siemens MAGNETOM Sykra MRI scanner (Siemens AG, Erlangen, Germany), with a 20-channel head array. Key presses were recorded using an MR-compatible response keypad (Current Designs, Philadelphia, USA). MR compatible headphones (Serene Sound Digital, Resonance Technology Inc., Los Angeles, USA) delivered acoustic stimuli.
During functional measurements, 1010 T2*-weighted gradient echo planar imaging (EPI) volumes (time of repetition (TR) = 1780 ms, time of echo (TE) = 30 ms, flip angle α = 90°, field of view (FoV) = 192 × 192 mm², voxel-size = 3.0 × 3.0 × 3.0 mm³) were obtained in one session. Volumes consisted of 32 slices (gap of 3.75 mm), covering the whole brain including cerebellum. Before the actual experimental session, a high-resolution structural volume was obtained covering the whole brain including cerebellum. Before the actual experiment, participants were de-randomized to ensure that the psychometric function started with more speech than non-speech features and receive feedback on their performance: a green emoji for correct answers, a red one for incorrect answers, a yellow one for the 50–50 stimuli, and in 25% of the trials, irrespectively of participants’ answer. Additionally, the red emoji was also presented if participants’ RT was slower than 1000 ms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### 2.6. Behavioral data analysis

#### 2.6.1. Performance

During the behavioral data analysis, the focus was first placed on the categorization performance of the participants. A calculation was made of the psychometric function, as the amount of non-speech in the stimuli as a function of each participant’s rating of the stimuli as non-speech (http://personalpages.manchester.ac.uk/staff/d.h.foster/software-modelfree/latest/index.html). Even though a trial-based performance feedback was provided, which allowed participants to learn the correct categorization, there were still participants who failed to perform the task correctly. Whether they failed because of task misunderstandings or because they were unable to differentiate the sounds remains unclear. The psychometric functions were used to divide participants into “good performers” and “poor performers”. Good performers were defined as those whose psychometric function started with <10% (discriminating a speech stimulus as non-speech) and ended with >90% (discriminating a non-speech stimulus as non-speech). This procedure resulted in 14 good performers and nine poor performers. After grouping participants, a t-test with the dependent variable Slope (at 50%) of the psychometric function was performed to validate differences between the groups.

#### 2.6.2. Reaction times

Reaction times (RT) shorter than 100 ms and longer than 2000 ms were excluded from further analysis (24 out of 4048 trials, 0.5%). The remaining data were analyzed in a 2 × 2 × 2 × 2 repeated measures Analyses of Variance (ANOVA) with the within-participant factors Correctness (correct vs. incorrect answers, independent of feedback), Time (first vs. second half of the experiment), and Speech (speech vs. non-speech), and the between-participant factor, Performance (good vs. poor). The 50–50 trials were not included in the analyses because it is not possible to assign them as either correct or incorrect. As expected, because they were only added to vary the task conditions, preliminary analyses showed no difference between the short and long stimuli, t(1, 22) = 1.03, p = .315. Therefore, this factor was combined.

### 2.7. FMRI data analysis

#### 2.7.1. Preprocessing

The FMRI data were processed and analyzed using SPM 12 (FIL, Wellcome Trust Centre for Neuroimaging, UCL, London, United Kingdom). To correct head motion, the functional time series were spatially realigned to the first image of the session. The structural T1-weighted volume was registered to the mean functional image and segmented in order to obtain spatial normalization parameters. Using these parameters, functional and structural images were normalized to the Montreal Neurological Institute (MNI) brain template. Finally, normalized functional volumes were smoothed with a three-dimensional Gaussian kernel of 8 mm full-width-half-maximum.

#### 2.7.2. First-level models

First-level models were built for each participant including both parts of the experiment: sound presentation and feedback presentation. The design of this study allowed for the analysis of those time points separately. Each single-participant model contained ten regressors of interest for the time point of sound presentation and six regressors of interest for the time point of the feedback presentation. For the sound presentation, eight regressors corresponded to the BOLD responses for correct and incorrect responses to either speech or non-speech stimuli in the first and second halves of the experiment, respectively. Further, BOLD responses for the 50–50 stimuli in both halves of the experiment were entered into two additional regressors. Since it is not possible to judge the categorization of those stimuli as either correct or incorrect, these were not considered in the second-level analyses. For feedback presentation, six regressors were built according to the BOLD responses for correct, incorrect, and neutral responses, representing the green, red, and yellow feedback conditions, for the first and second halves of the experiment.

Additionally, signal changes related to head movement were accounted for by including the six movement parameters calculated in the SPM12 realignment procedure as additional regressors. Time series in each voxel were high-pass filtered to 1/128 Hz and modeled for temporal autocorrelation across scans with an autoregressive (AR (1)) process. On the basis of the first-level model, contrasts for the second-level model were calculated. Two separate full-factorial ANOVA designs were used, one for the sound presentation and one for the feedback presentation.
2.7.3. Second-Level Model: Sound presentation

The single-participant contrasts for the different conditions (excluding the 50–50 condition) were entered into a $2 \times 2 \times 2 \times 2$ full-factorial ANOVA with the within-participant factors Speech (speech vs. non-speech), Correctness (correct vs. incorrect answers, independent of feedback), and Time (first vs. second half of the experiment) and the between-participant factor Performance (good vs. poor). Within this ANOVA, F-contrasts for all main effects and interactions were calculated. In order to reveal the underlying direction of the main effects, significant activations in F-contrasts were followed by t-contrasts.

2.7.4. Second-Level Model: Feedback presentation

Regarding feedback presentation, the focus was placed on the processing of the different feedback types. Therefore, the single-participant contrasts for the different feedback conditions were entered into a $3 \times 2 \times 2 \times 2$ full-factorial ANOVA with the within-participant factors Feedback (correct, incorrect, neutral), Time (first vs. second half of the experiment) and the between-participant factor Performance (good vs. poor). Within this ANOVA, F-contrasts for all main effects and interactions were calculated. In order to reveal the underlying direction of the main effects, significant activations in F-contrasts were followed by t-contrasts.

2.7.5. Correction

Results of all analyses are reported as statistically significant at a threshold of $p < .05$, corrected for whole brain with family-wise errors (FWE) and a cluster size $k = 20$ or corrected for region of interest in Heschl’s gyrus and superior temporal gyrus (as included in the WFU PickAtlas extension for SPM, $p < .001$, uncorrected).

In order to further visualize the significant effects of feedback presentation, average beta values as a function of feedback (green, red, and yellow) in a sphere with a radius of 8 mm around the activation peak maxima were extracted. This type of visualization does not contain circularity effects according to Kriegeskorte, Simmons, Bellgowan, and Baker (2009), since an ANOVA was used and the source of significance within a main effect or interaction was subsequently determined. This approach is an extension of the same analyses and not double dipping. Note that the extraction of beta values was only illustrative, and inferences were taken from original analyses.

3. Results

3.1. Behavioral data

3.1.1. Performance

The results of the t-test with the factor Slope of the psychometric function showed a significant difference between both groups, $t (21) = -3.69, p = .001$, indicating a steeper psychometric function for good performers compared to poor performers. This result further underlined our decision to analyze both groups separately. Note that slope cannot be entered as a covariate into the ANOVA because there is too much variance in the group of poor performers. There are steeper slopes, but participants may not fit into the inclusion criteria described above since their performance on the easier trials was worse (see Fig. 3). Therefore, a between-participant ANOVA was performed.

3.1.2. Reaction times dependent on performance

The results of the behavioral data are shown in Fig. 4, mean RTs for the different conditions are shown in Table 1. Analyses of the RT data revealed a main effect of Correctness, $F(1, 21) = 12.73, p = .002$, $\eta^2_p = .377$, with faster responses for correct (703 ms) compared to incorrect (716 ms) answers, and a main effect of Speech, $F(1, 21) = 5.43, p = .030$, $\eta^2_p = .205$, with faster responses for speech (716 ms) compared to non-speech stimuli (742 ms). None of the other main effects approached significance. The two-way interaction Time x Correctness, $F (1, 21) = 6.65, p = .018$, $\eta^2_p = .240$, and the three-way interaction Time x Correctness x Performance, $F(1, 21) = 7.40, p = .013$, $\eta^2_p = .260$ were significant. The two-way Time x Correctness interaction indicated that, over time, the RT for correct responses did not differ (first half: 704 ms vs. second half: 703 ms), whereas incorrect answers got slower toward the end of the experiment (777 ms) compared to the beginning (731 ms). This interaction is further supported by the three-way interaction Time x Correctness x Performance pointing toward group differences regarding this effect. Further post hoc analyses showed a significant interaction Time x Correctness, $F(1, 13) = 13.22, p = .003$, $\eta^2_p = .504$ only in the group of good performers. Furthermore, there was a tendency toward the interaction Correctness x Performance, $F (1, 21) = 3.68, p = .069$, $\eta^2_p = .149$, indicating that in the group of good performers, there were large differences in RTs between correct (665 ms) and incorrect (743 ms) responses, whereas in the group of poor performers, there was almost no difference (correct: 742 ms vs. incorrect: 765 ms). No other main effects or interactions approached significance.

3.2. FMRI data

3.2.1. Sound presentation

As already indicated in the behavioral data, performance differences between participants were also evident in the fMRI data. During sound presentation, regions were found that showed stronger activation among good performers compared to poor performers and regions that showed a contrary pattern (see Table 2A, B). Stronger activation in good performers is seen for example in the left auditory cortex as well as the right superior and middle temporal gyrus, the caudate nucleus, and the lingual gyrus; whereas a stronger activation for poor performers is mainly seen in the bilateral precuneus, the anterior cingulate cortex, and the supplemental motor area, but also in the superior middle temporal gyrus. Using region of interest analysis in the bilateral Heschl’s gyrus and the superior temporal gyrus, different activations were found in both groups (see Fig. 5).

In line with the behavioral and fMRI data, there was a main effect of Time, with higher neuronal activity in the first compared to the second half of the experiment in the bilateral post- and precentral gyrus, the dopaminergic midbrain areas (such as pallidum and caudate nucleus), and the frontal regions (see Table 2C). None of the other main effects and interactions showed any significant results.

3.2.2. Feedback presentation

In line with the results during sound presentation, neuronal data during feedback presentation also revealed a main effect of Performance in both directions. Good performers showed stronger activation in the bilateral middle temporal and the occipital gyrus whereas poor performers showed stronger activation in the frontal regions as well as the calcarine and precuneus (see Table 3A, B).

In addition to the main effect of Performance, there was also a main effect of Time with higher neural activity in the first compared to the second half of the experiment, mainly in frontal and temporal regions (see Table 3C).

Activations regarding the main effect of Feedback (see Table 3D) were found in bilateral frontal regions, the supplemental motor area, and the middle cingulate cortex, but also in the bilateral precuneus and middle temporal gyrus. A closer evaluation of this main effect, by extracting the mean beta values in each peak activation, revealed two different patterns for the different levels. On the one hand, there was a positive effect for all feedback types, indicating stronger activation for negative compared to positive feedback in the frontal region and the angular gyrus (see Fig. 6, green). On the other hand, there was a stronger deactivation for negative feedback than for positive feedback in the middle cingulate cortex, supplemental motor area, insular cortex, and medial temporal gyrus (see Fig. 6, red). There was no significant difference with the neutral feedback in any of the regions that survived the FWE correction.
4. Discussion

The goal of the current study was to investigate processing differences between blended speech and non-speech for which categorization was more difficult. A two-alternative forced choice task was used, which included feedback. Therefore, the difference in processing of positive, negative, and neutral outcomes was also investigated.

4.1. Performance

Despite feedback, results of the behavioral data, especially the psychometric functions, indicated that there were some participants who were never able to solve the task, even when presented with easier conditions. Using the individual psychometric functions over different speech-non-speech levels, the participants were divided into a group of good performers, those who learned the categorization of the stimuli into speech and non-speech sounds, and a group of poor performers, those who had problems differentiating the stimuli even in the easier conditions. The reason for including the performance feedback, was to allow participants to learn the task and to prevent misunderstandings of task instructions. Still, the task was difficult for some participants to solve. One reason may be the similar complexity of speech and non-

![Fig. 3. Fitted individual and mean psychometric function for the good and poor performance group separately. The percentage of non-speech responses given by the participants were plotted against the percentage of non-speech contained in the stimulus.](image)

![Fig. 4. Mean RT [ms] as functions of Speech (speech vs. non-speech, left vs. right column), Time (first vs. second half of the experiment) and Correctness (correct in dark grey vs. false in light grey) are shown for good (upper panel) and poor performers (lower panel).](image)
speech sounds in the present study, an advantage in contrast to studies morphing speech sounds with white noise (Osnes, Hugdahl, Hjelmervik, & Specht, et al., 2011; Osnes, Hugdahl, & Specht, 2011; Specht et al., 2011). Therefore, only pure speech stimuli were categorized as speech and whenever there was a change in the stimulus, it was categorized as non-speech. This did not change over the time course of the experiment. There was no difference between correct and incorrect responses in their reaction times, since they did not learn the categorization. The task difficulty, induced by the different levels of speech-non-speech stimuli, together with the negative feedback that was given for responses longer than 1000 ms, inevitably led to an increased amount of negative feedback. Hence, this negative feedback may have increased uncertainty about the categorization in poor performing participants. Consequently, despite the feedback, these participants would not have been able to learn the categorization. Moreover, the larger amount of negative feedback could also have reduced motivation.

4.2. Speech – non-speech processing

It is known that speech processing is organized hierarchically in a dorsal and a ventral pathway (Rauschecker & Scott, 2009). The dorsal path is responsible for spatial processing and includes bilateral auditory areas, while the ventral path is able to identify complex auditory patterns, such as speech and non-speech, and is mainly lateralized on the left hemisphere (Liebenthal et al., 2005; Narain et al., 2003). Specifically, this pathway includes more of the anterior areas in the superior temporal gyrus and appears to be sensitive to spectro-temporal complexity and linguistic intelligibility (Binder et al., 2000; Giraud et al., 2004; Narain et al., 2003; Poeppel, 2003; Scott et al., 2000).

There was no overall differences found between speech and non-speech stimuli (separated at the 50–50 stimulus), which might be due to the blended stimuli. Almost every stimulus contained parts of speech and parts of non-speech and was therefore not unambiguously associated with one category or the other. However, there were differences in neural activation in the auditory areas between the good and poor performers. Results in the different groups showed a left lateralization in the good performance group with activations in the primary auditory cortex areas and anterior superior temporal sulcus. The latter region was shown to be involved in processing intelligible speech stimuli by Scott et al. (2000), and indicates that, participants were indeed able to identify the “speech” parts in the stimulus, even when intermixed with between correct and incorrect responses over the time course of the experiment for the good performance group, whereas there was no difference in the other group. During the first part of the experiment, good performers learned to categorize the stimuli according to the criterion. Consequently, in the second part, their responses were fast and correct in most of the trials, and thus they received positive feedback. For those trials in which they were still uncertain about the stimulus category, they needed more time for their decision resulting in a larger amount of incorrect responses in the difficult trials. The poor performers, in contrast, tended to categorize stimuli as non-speech stimuli even if there was only a slight amount of non-speech included, and despite the feedback that could have guided them towards the desired decision criterion. One reason for the poor performance might be that for those participants, familiar sounds, such as vowels, seem to be perceived in a more categorical way compared to the non-speech sounds, which the participants had not experienced before (Elmas, 1975; Goto, 1971; Werker & Tessen, 1984). Therefore, only pure speech stimuli were categorized as speech and whenever there was a change in the stimulus, it was categorized as non-speech. This did not change over the time course of the experiment. There was no difference between correct and incorrect responses in their reaction times, since they did not learn the categorization. The task difficulty, induced by the different levels of speech-non-speech stimuli, together with the negative feedback that was given for responses longer than 1000 ms, inevitably led to an increased amount of negative feedback. Hence, this negative feedback may have increased uncertainty about the categorization in poor performing participants. Consequently, despite the feedback, these participants would not have been able to learn the categorization. Moreover, the larger amount of negative feedback could also have reduced motivation.
non-speech parts. In the study by Scott et al. (2000), noise vocoded or spectrally rotated speech resulted in activation in the bilateral superior temporal gyrus. This bilateral pattern of activation is seen in our group of poor performers, indicating that, for them, most of the stimuli sounded like non-speech, and were processed accordingly. In addition to the anterior superior temporal gyrus, the ventral pathway also shows projections toward the inferior frontal areas (Hackett, Stepniewska, & Kaas, 1999; Romanski et al., 1999), which was found in this data as well, with stronger activation in the good performance group. Thus, both activation patterns underline the assumption that in the good performance group, participants were able to discriminate stimuli and consequently activate the ventral pathway for processing of speech sounds, whereas the poor performance group failed to discriminate the stimuli and thus used the dorsal pathway as well. The fMRI results are in line with the behavioral results and underline the assumption that the poor performance group was not able to learn the discrimination, since they used a different processing strategy compared to the good performance group. This pattern of results speaks against a general misunderstanding of task instructions by some individuals.

Even though performance feedback was provided, 9 out of 23 participants still failed to perform the task correctly. Since speech can be understood even in a variety of disturbing conditions (e.g., Holt & Lotto, 2010; Liberman et al., 1967; Liberman, 1996; Liberman & Mattingly, 1989), it could be concluded that the task is not related to speech perception. However, the results of the fMRI data disconfirm such an interpretation, because speech related areas were found only in the good performing group, whereas there was a bilateral activation in the poor performing group. If the task was not related to speech perception, there would not have been the speech related areas observed in the good performing group. This activation, in conjunction with the feedback related activations discussed below and the results of the behavioral task, provide evidence that the task is related to speech perception. The poor performing participants are indeed able to identify the pure speech sounds, but were not able to differentiate between the various kinds of blended stimuli. Despite feedback, however, they might not be able to apply the correct decision criterion, because of perception or motivational reasons (e.g. many negative feedback trials).

As there was also an interest in feedback processing, non-informative trials (neutral feedback) as well as a time-out criterion (RTs larger than one second, resulting in negative feedback, irrespective of the correctness of the response) were included. Those feedback conditions might have disturbed the performance of the participants who already were uncertain about the categorization. The increased amount of negative feedback may have also resulted in less motivation. Since most of the participants were able to discriminate the stimuli, and behavioral results are in line with the neuronal data, it is argued that our task is related to speech perception.

### 4.3. Feedback anticipation

Although the direct comparison between correct and incorrect responses did not result in any significant effect after whole-brain FWE correction, a different contrast can be associated with the feedback anticipation in the current experiment: comparison between the first and second halves of the experiment. It can be assumed that participants learned to discriminate speech and non-speech stimuli during the

---

**Table 3**

<table>
<thead>
<tr>
<th>Contrast</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>Volume</th>
<th>Z</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(A) Performance: Good &gt; Poor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>−74</td>
<td>−4</td>
<td>299</td>
<td>6.36</td>
<td>Middle temporal/occipital gyrus R (BA37)</td>
<td></td>
</tr>
<tr>
<td>−40</td>
<td>−76</td>
<td>24</td>
<td>60</td>
<td>5.95</td>
<td>Middle temporal/occipital gyrus L</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>−14</td>
<td>−24</td>
<td>24</td>
<td>5.75</td>
<td>Superior/middle temporal gyrus R</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>−52</td>
<td>−18</td>
<td>22</td>
<td>5.45</td>
<td>Fusiform gyrus R (BA37)</td>
<td></td>
</tr>
<tr>
<td>−44</td>
<td>−84</td>
<td>10</td>
<td>40</td>
<td>5.35</td>
<td>Middle occipital gyrus L</td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>48</td>
<td>−30</td>
<td>−4</td>
<td>46</td>
<td>4.59</td>
<td>Middle temporal gyrus R</td>
</tr>
<tr>
<td><strong>(B) Performance: Poor &gt; Good</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>−14</td>
<td>−62</td>
<td>30</td>
<td>122</td>
<td>6.11</td>
<td>Precuneus L (BA31), Cuneus L</td>
<td></td>
</tr>
<tr>
<td>−34</td>
<td>36</td>
<td>−8</td>
<td>52</td>
<td>5.73</td>
<td>Middle frontal gyrus L (BA47)</td>
<td></td>
</tr>
<tr>
<td>−6</td>
<td>56</td>
<td>−8</td>
<td>55</td>
<td>5.40</td>
<td>Medial frontal gyrus L (BA10)</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>−64</td>
<td>8</td>
<td>20</td>
<td>5.36</td>
<td>Calcarine R</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>−64</td>
<td>4</td>
<td>48</td>
<td>5.35</td>
<td>Superior frontal gyrus R</td>
<td></td>
</tr>
<tr>
<td><strong>(C) Time: 1st &gt; 2nd</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>46</td>
<td>36</td>
<td>29</td>
<td>5.62</td>
<td>Superior frontal gyrus R</td>
<td></td>
</tr>
<tr>
<td>−2</td>
<td>32</td>
<td>54</td>
<td>51</td>
<td>5.35</td>
<td>Superior frontal gyrus L</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>24</td>
<td>10</td>
<td>19</td>
<td>5.26</td>
<td>Inferior frontal gyrus R</td>
<td></td>
</tr>
<tr>
<td>−48</td>
<td>16</td>
<td>28</td>
<td>52</td>
<td>5.14</td>
<td>Inferior frontal gyrus L</td>
<td></td>
</tr>
<tr>
<td>−2</td>
<td>40</td>
<td>38</td>
<td>97</td>
<td>5.13</td>
<td>Medial frontal gyrus L</td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>68</td>
<td>−42</td>
<td>14</td>
<td>4.82</td>
<td>Superior temporal gyrus R</td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>54</td>
<td>−8</td>
<td>−14</td>
<td>19</td>
<td>3.63</td>
<td>Middle temporal gyrus R</td>
</tr>
<tr>
<td><strong>(D) Feedback</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>36</td>
<td>92</td>
<td>6.25</td>
<td>Middle cingulate cortex L/R</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>62</td>
<td>116</td>
<td>5.79</td>
<td>Supplemental motor area R</td>
<td></td>
</tr>
<tr>
<td>−6</td>
<td>54</td>
<td>−8</td>
<td>44</td>
<td>5.46</td>
<td>Medial frontal gyrus L</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>22</td>
<td>−12</td>
<td>37</td>
<td>5.53</td>
<td>Insula R/inferior frontal gyrus R</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>54</td>
<td>18</td>
<td>20</td>
<td>5.21</td>
<td>Superior frontal gyrus R</td>
<td></td>
</tr>
<tr>
<td>−20</td>
<td>28</td>
<td>48</td>
<td>55</td>
<td>5.51</td>
<td>Superior frontal gyrus L</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>−70</td>
<td>32</td>
<td>61</td>
<td>5.25</td>
<td>Angular gyrus L</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>−24</td>
<td>−6</td>
<td>22</td>
<td>5.44</td>
<td>Middle temporal gyrus R (*)</td>
<td></td>
</tr>
<tr>
<td>−6</td>
<td>−54</td>
<td>20</td>
<td>101</td>
<td>5.69</td>
<td>Precuneus L/R</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5. FMRI results during sound presentation showing the main effect of Performance with higher activity for good performers in white and higher activity for poor performers in black. Activations are superimposed on the mean of the individual participant’s T1 images (p < .05, k > 20 (FWE corrected)).
first half of the experiment, at least in the good performance group. In the second half of the experiment, the discrimination criterion between speech and non-speech stimuli is settled. In the good performance group, the only sources of errors were those stimuli around the 50–50 stimulus, which are almost impossible to discriminate. When comparing the first and second halves of the experiment, there was higher neural activity, mainly in the left caudate and pallidum and the bilateral auditory areas. In the first half of the experiment, a stronger activation was expected in those areas in correct trials compared to incorrect trials (Puschmann et al., 2013; Thiel, Bentley, et al., 2002; Weis et al., 2012). This pattern reveals that learning speech-non-speech discrimination occurred mainly in the first half of the experiment, whereas feedback did not play a major role in the second half of the experiment, where the discrimination criterion is established. Stark and Scheich (1997) showed that the dopaminergic system is only activated in early phases of learning, whereas Reed et al. (2011) showed unstable learning related changes in the auditory cortex that fade over time.

The performance of the participants also influenced the anticipation of the different feedback types. In the group of good performers, activation occurred in the left caudate nucleus, which is known to be a part of the reward system, and thus indicated feedback, especially reward anticipation. In contrast, the poor performance group, who also erroneously judged more stimuli as non-speech, showed enhanced activation in the supplemental motor area, anterior cingulate cortex, angular gyrus, and insular cortex. Those regions are known to be involved in error processing, notation of contradictions, and the evaluation of outcomes (e.g., Holroyd et al., 2004; Holroyd & Yeung, 2012; Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006; Scheffers, Coles, Bernstein, Gehring, & Donchin, 1996; Taylor, Stern, & Gehring, 2007). This suggests that they tried to solve the task correctly but had problems in processing the blended stimuli and identify the speech within the stimuli.

4.4. Feedback processing

During feedback delivery, it was hypothesized that some activity would be found in the auditory cortex areas related to the different feedback types, as was seen in some previous studies (Pleger et al., 2007; Weil et al., 2010; Weis, Brechmann, et al., 2013; Weis, Puschmann, et al., 2013). In contrast to these previous studies, the present study found a cluster of activation in the right middle temporal gyrus (almost the same region in which it was found by Weis, Puschmann, et al., 2013), with stronger activation for negative feedback than positive feedback. This difference compared to the previous studies, where a stronger activation was found in the left and right auditory cortices depending on the learning status of the participants (Weis, Brechmann, et al., 2013), can be explained by the fact that, in the current study, an informative reward rather than a monetary reward was used. Reward-related plasticity in the auditory cortex, as was found in the previous studies, with stronger activation when the expectation of the participants was correct, appears to occur only when the feedback works as a reward. This might not be the case when using only informative feedback. To underline this explanation, in the feedback contrast, no activation in reward-related areas was found. Thus, the feedback used in the current study was not as effective as the monetary reward in the previous studies.

The good performance group showed stronger activation in the right middle temporal gyrus, close to the fusiform gyrus (BA37) in the fusiform face area (Rossion, Schiltz, & Crommelinck, 2003), indicating a more detailed processing of the emojis compared to the poor performance group, who received more red than green emojis. In contrast, the latter group showed stronger activation in the left precuneus (BA31) as well as the left frontal regions (BA47, BA10), both known to be involved in source memory (Lundstrom, Ingvar, & Petersson, 2005). Participants in the poor performance group still attempted to solve the task correctly by remembering the sounds and evaluating the feedback.

Comparably to the findings during the sound presentation, there was an effect of time, with stronger activation in the bilateral superior and the inferior frontal regions during the first half of the study compared to the second half. Those regions mainly comprise BA9, which is known to be involved in short-term memory (Babiloni et al., 2005), i.e., remembering the previous stimulus and evaluating the outcome, including error detection (Chevrier, Noseworthy, & Schachar, 2007) and auditory verbal attention (Nakai, Kato, & Matsuo, 2005). This indicates that, together with the findings during the sound presentation, participants in the first half learned to discriminate between the speech and
non-speech stimuli, whereas in the second half of the experiment, the discrimination pattern is established.

4.5. Main effect of feedback

The main effect of feedback involves several regions, such as the middle cingulate cortex, supplemental motor area, bilateral frontal gyrus, bilateral precuneus, and right middle temporal gyrus. However, the factor feedback contains three different levels: positive, negative, and neutral. Therefore, extraction of the beta values gives us the possibility of investigating the underlying pattern of activation. There are two main results: (1) regions with a deactivation that were mostly left lateralized, which is strongest for negative, followed by neutral and positive feedback, such as the left medial and superior frontal gyrus, angular gyrus, and precuneus, as part of the default mode network (Frings et al., 2010; Raichle et al., 2001; Seghier, Fagan, & Price, 2010); and (2) regions with the strongest activation for negative, followed by neutral and positive feedback were mostly right lateralized, such as the supplemental motor area, bilateral insular cortex, and middle cingulate cortex, but also the middle temporal gyrus. Results in the supplemental motor area and bilateral insula cortex were linked to error processing, as shown in many EEG studies, as the supplemental motor area is involved in error-related negativity (Holroyd et al., 2004; Scheffers et al., 1996; Taylor et al., 2007). Previous fMRI studies have also reported stronger activation for negative feedback compared to positive feedback within the supplemental motor area (Özyurt, Rietze, & Thiel, 2012; Ullsperger and von Crumon, 2004).

There were no significant differences between positive and neutral feedback or negative and neutral feedback in any of the regions. In other words, neutral feedback works as an intermediate feedback. There was no activation found in the primary auditory cortex due to the differences in the expectations of participants, which would have shown stronger activation for both correct and incorrect feedback than for neutral feedback.

4.6. Limitations

In contrast to the hypothesis, there was no significant activation found in auditory cortex areas related to feedback presentation. It was expected that there would have been a stronger activation for correct and incorrect responses compared to neutral responses. The results showed that at least some participants were able to correctly differentiate both stimulus types after a learning phase. In the case of neutral feedback, there should be less activation in the auditory cortex, since participants’ expectations (correct or incorrect feedback) were not fulfilled. However, the paradigm of this study may have failed for the following reasons.

First, in contrast to the previous studies, informative feedback was used instead of a monetary reward. This could also explain the motivation observed in the poor performing group to solve the task correctly. As discussed above, some results indicated that the reward did not work as expected in the current experiment (i.e., only a small activation in reward-related areas during reward anticipation, and no difference in reward-related areas during reward delivery). Most of the previous studies that have investigated the effect of feedback anticipation used a monetary reward (Gehring & Willoughby, 2002; Kirsch et al., 2003; Knutson, Westdorp, Kaiser, & Hommer, 2000; Knutson, Adams, Fong, & Hommer, 2001; Knutson et al., 2000; Pleger et al., 2008; Weil et al., 2010; Weis et al., 2012; Weis, Puschmann, et al., 2013), virtual winning or losing in gaming tasks (Cohen & Ranganath, 2007), or liquid rewards as used in some studies in animals (Beitel et al., 2003; Blake et al., 2006; Brosch, Seleznева, & Scheich, 2011) and humans (O’Doherty, Deichmann, Critchley, & Dolan, 2002).

Second, some of the participants were not able to solve the task accordingly. For them, most of the stimuli sounded like non-speech, even those that contained a significant amount of speech material. Consequently, participants had to be separated into good and poor performers, and both groups showed different patterns of activation during sound and feedback presentation. If all participants had categorized the stimuli correctly, the results may be clearer.

4.7. Conclusion

In summary, this study demonstrated that, even when speech and non-speech stimuli are blended together, most participants were able to differentiate the stimuli correctly. This allowed us to determine the speech-related areas in the good performance group, even if there was no main effect of speech. Categorization of blended speech and non-speech stimuli is not the same for all participants. Some of them judged the stimuli according to the established criterion, but there was also a group that judged almost all stimuli as non-speech even if there was a larger amount of speech included in the stimuli. Furthermore, it was revealed that when it comes to two-alternative forced choice paradigms, informative feedback might not be as successful as a monetary reward, as there was no reactivation of the sensory cortices as expected. In general, the results found during sound presentation are in line with those found during the feedback presentation. Thus, it was shown that the previously used paradigms work with a different class of stimuli, namely speech and non-speech stimuli, and that participants were able to discriminate the stimuli. Consequently, results may reflect how well humans can estimate the percentage of speech in a signal. Future studies should improve the paradigm by ensuring that more participants are able to categorize the stimuli correctly and by using feedback that is more rewarding than an informative green or red emoji.

References


