Atten Percept Psychophys (2016) 78:1496–1511
DOI 10.3758/s13414-016-1087-6

# Speech imagery recalibrates speech-perception boundaries

**Mark Scott**[1,2]

Published online: 11 April 2016


**Abstract** The perceptual boundaries between speech sounds are malleable and can shift after repeated exposure to contextual information. This shift is known as *recalibration*. To date, the known inducers of recalibration are lexical (including phonotactic) information, lip-read information and reading. The experiments reported here are a proof-of-effect demonstration that speech imagery can also induce recalibration.

**Keywords** Recalibration · Speech imagery · Speech perception · Sensory adaptation · Forward models

## Introduction

The speech-perception system is capable of rapidly adjusting its perceptual boundaries. This adjustment, or *recalibration*, is useful for accommodating to differences between speakers' productions of speech sounds — for example making it easier to understand the locals when moving to a region where a different dialect is spoken, or to understand the speech of someone with a speech impediment. In order for recalibration to occur, the perceiver necessarily needs to learn that an atypical instance of a speech sound belongs to a particular category. If the sound is atypical, though, how does the perceiver know which category it belongs to? It could be that the correct categorization is arrived at via lexical knowledge: If only one of the possible candidate categories of an atypical sound would produce a real word, then that is good evidence for the category.[1] Similarly, visual information can help determine the intended category: If the speaker's face is clearly pronouncing a sound involving lip closure, then the candidate /b/ is far more likely than /d/. In addition, reading can indicate the correct categorization of an ambiguous sound. All of these causes (lexical, visual, reading) have been shown to induce recalibration. The current experiment tests whether the influence of speech imagery (the 'voice in one's head') can also induce recalibration.

The possibility that speech imagery can induce recalibration is motivated in part by the recent discovery that speech imagery can alter speech perception (Scott et al., 2013).[2] When participants are asked to imagine a speech sound in time to an ambiguous external speech sound, they tend to hear the ambiguous sound as matching the content of their imagery (Scott et al., 2013). This is similar to the well-known visual influence on speech perception, in which video of a face pronouncing a speech sound can alter the auditory perception of speech — e.g., the "McGurk effect" (McGurk and MacDonald, 1976), or "visual dominance" as found in Rosenblum and Saldaña (1992). In both cases,

This research was supported by start-up research grant 31h060 from UAEU

✉ Mark Scott
mark.a.j.scott@gmail.com

[1] United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

[2] Present address: Qatar University, Doha, Qatar

[1] Similarly, recalibration occurs if only one of the words is a *possible* word — i.e. based on phonotactic information (Cutler et al., 2008).

[2] Sams et al. (2005) performed a similar experiment showing the influence on perception of moving the speech articulators. Sams' et al. experiment was not about speech imagery and so did not discuss the connection between this movement of the articulators and speech imagery, nor did it test "pure" speech imagery (without movement of the articulators).

there is an auditory speech stimulus whose perception is pushed into alignment with the category indicated in a separate information source (vision or imagery). As the visual effect on speech perception can induce recalibration, this leads to the question of whether speech imagery can have a similar effect.

Another motivation for the current study is the proposed mechanism of inner speech, which has been theorized to be produced by the motor system by means of *forward models* (see Pickering and Garrod (2013) and Scott (2013a), and for a related discussion, Tian and Poeppel (2010)). These forward models are discussed in detail in "Forward models", but in essence they are a component of the motor system that predicts self-caused sensations. The theory is that these sensory predictions have been co-opted to provide the sensory experience of speech imagery. These forward models have also been implicated in disambiguation of ambiguous speech sounds (Pickering & Garrod, 2007) which is a key aspect of recalibration. Similarly, visual influences on speech perception have been theorized to be mediated by forward models (Skipper et al., 2007).

To be clear, the purpose of these experiments is to provide evidence that speech imagery can induce recalibration. The fact that speech imagery can alter perception and the fact that speech imagery is apparently tied to forward models, which may also underlie other forms of recalibration, are motivations for the current set of experiments. However, these experiments are not intended to test the mechanisms of recalibration, they are intended solely as a "proof of effect".

## Speech imagery

Speech imagery (or "inner speech") is a ubiquitous, though often overlooked, mental phenomenon, occupying perhaps a quarter of our conscious time (Heavey & Hurlburt, 2008) and linked to many crucial aspects of cognition.

Inner speech comes in at least two distinct forms (both tested in the experiments reported here), one can 'mouth' words while talking to oneself (*enacted* inner speech), or one can keep the articulators immobile (*non-enacted* or *pure* inner speech). In both cases, one 'hears' one's voice internally without any audible external sound being produced. Oppenheim and Dell (2010) have argued that inner speech shows a 'flexible abstractness', meaning that the degree of phonetic detail present in our experience of inner speech is dependent on how much motor engagement there is. When there is no motor engagement, inner speech is a purely abstract phonological code, but when the motor system is engaged (in *mouthing*), inner speech contains more phonetic detail.

Recent research has suggested that both forms of imagery are tied to *forward models* in the motor system (discussed in "Forward models"). The evidence for this connection is both behavioural (Scott, 2013a) and from brain imaging (Tian & Poeppel, 2010).

## Recalibration

Several experiments have demonstrated recalibration from both visual and lexical influences (for a good review of the relevant literature, see Samuel and Kraljic, 2009), and very recently a study has also shown recalibration from reading (Keetels et al., 2016).

That lexical information can influence speech perception was first shown by Ganong (1980), who demonstrated that people are more likely to hear an ambiguous speech sound as belonging to the category that allows the perceived word to be a real word. This "Ganong" effect can induce recalibration, as was first demonstrated by Norris et al. (2003). In this experiment, listeners were repeatedly exposed to the Ganong effect, inducing them to hear a sound as either /f/ or /s/, after which they categorized sounds from an /f/ to /s/ continuum. Those who had been induced to hear /f/ in the exposure session continued to categorize the sounds as /f/ in the test session, and contrariwise for those exposed to /s/. This effect has been replicated many times (e.g., Eisner and McQueen 2005; Kraljic and Samuel 2005; Kraljic et al. 2008a; Sjerps and McQueen 2010).

It is not only the lexical status of a word that influences perception in this way, but its *possible* lexical status. Cutler et al. (2008) showed that recalibration occurs when only one of the possible candidate sounds is phonotactically possible. It is unclear whether this should be considered a special case of lexically-driven recalibration or a separate source of recalibration. For the moment, I will assume phonotactic recalibration is a special case of lexically-induced recalibration, though nothing in this paper depends on the distinction.

Recalibration from speech-read (or 'lip-read') visual information was first reported by Bertelson et al. (2003). In that experiment they showed that when video of a face pronouncing /aba/ or /ada/ was matched with audio that was ambiguous between these two sounds, participants heard the sound as belonging to the category indicated by the video. That much is simply the well-established influence of vision on speech perception (e.g., McGurk and MacDonald, 1976), but interestingly, after repeated exposure to one of these 'visually shifted' stimuli, participants experienced an aftereffect — when the video was removed, they continued to categorize the ambiguous sound as they had when the video

was present. Their phoneme categories had *recalibrated*. This effect has also been replicated many times (e.g., van Linden and Vroomen 2008; Vroomen et al. 2004; Vroomen and Baart 2009a).

Both techniques induce recalibration, though recalibration due to lexical information appears to last longer (Kraljic and Samuel, 2005; Eisner & McQueen, 2006) than that induced by visual information (Vroomen & Baart, 2009b). Furthermore, Reinisch et al. (2014) report that it is a widely held, though recently challenged, assumption that the means by which disambiguation occurs – lexical or visual – is largely irrelevant. The assumption here is that recalibration is a matter of associating an ambiguous sound with a category and as long as the category is determined, the contextual information which makes the determination is largely irrelevant. This would suggest that simply telling people the category of an ambiguous sound may be sufficient to induce recalibration, which is taken up in **Experiment Three**.

A very recent study has also shown recalibration from merely reading text (Keetels et al., 2016). This experiment showed that getting participants to read a target word just before hearing the ambiguous sound in the exposure phase is sufficient to induce recalibration.

The mirror effect of recalibration is *Adaptation*. In an adaptation paradigm, participants are repeatedly exposed to a sound but, unlike in the recalibration paradigm, the sound is a clear, unambiguous representative of its category. After these repeated exposures, ambiguous sounds are less likely to be perceived as belonging to the category to which participants were just exposed. For example, Eimas and Corbit (1973) showed that after repeated exposure to /ba/, participants categorized *fewer* sounds from a /ba/ to /pa/ continuum as being /ba/ (and vice versa after exposure to /pa/).

There has already been some work looking at whether speech imagery can induce adaptation. However this work has been inconclusive, with an early study (Cooper et al., 1976) showing a small effect, but later studies failing to replicate this finding (e.g., Summerfield et al., 1980). A study in which participants *mouthed* along with a clear token failed to show any influence of speech imagery on levels of adaptation (Scott, 2013b). The situation with visual information is more clear: Visual information does not appear to induce adaptation (Saldaña & Rosenblum, 1994).

Based on current understanding there is no clear reason to predict that speech imagery should induce adaptation and the current experiments are aimed at examining recalibration, not adaptation.

## Forward models

As discussed in "Introduction", one motivation for this experiment is the theory that *forward models* underlie both the effects of inner speech on speech perception (Tian & Poeppel, 2010; Scott, 2013a) and the effects of visual information on speech perception (Skipper et al., 2006; Skipper et al., 2007). This is further tied to recalibration by the hypothesized role of forward models in disambiguating unclear speech sounds (Pickering & Garrod, 2007).[3]

A forward model is a component of the motor system which predicts the sensory consequences of one's own actions. This is a vital function as it allows for *ersatz* sensory feedback to guide actions in situations in which the action would be complete before *real* feedback could be transduced and processed — as with the very fast movements of the articulators during speech production (Miall & Wolpert, 1996). Forward models also allow for the 'tagging' of self-generated sensations so that we do not confuse the sensations we produce with those produced by something in the external world.

Several recent theories have proposed a role for forward models in speech perception. For example, (Pickering & Garrod, 2007) argue that when processing ambiguous sounds, predictions from the forward models can 'fill in' information that is unclear in the raw signal. Skipper et al. (2006, 2007) have proposed something similar.

These recent theories are very similar to the much-debated *Motor Theory of Speech Perception* (Liberman & Mattingly, 1985), which proposes that speech perception is inherently dependent on the motor-system, perceiving sounds in terms of the gestures that produced them. The primary difference between these recent *forward model* theories and the *motor theory* is that the forward model theories view the involvement of the motor system as a potential strategy to aid the perceptual system when the auditory signal is ambiguous. These theories do not claim that perception is necessarily achieved by means of the motor system. Nor do they claim that recovering the articulatory gesture is necessarily the goal of speech perception.

If these forward-model theories are correct, then when we hear an ambiguous speech sound, we consult our own motor systems (using forward models) to determine the

[3] It should be noted that the possibility that these different forms of recalibration are mediated by the perceiver's motor system does not imply that they should show equivalent levels of recalibration. The pathway to that mechanism is different in these cases which introduces extraneous factors that would make any prediction of equivalence moot.

intended word. Thus, in the case of lexical (or phonotactic) recalibration, the motor system is 'filling in' the missing information based on what is an already known sound sequence (lexical) or a possible sound sequence (phonotactics).

Forward models may also play a role in recalibration from visual information. The effect of visual processing on auditory speech perception has recently been theorized to involve forward models. The idea is that seeing a speech gesture triggers one's own forward models of that gesture and thus influences perception in a 'top-down' fashion (Skipper et al., 2007).

The possible role of forward models in recalibration motivates the investigation of imagery as a possible inducer of recalibration since several experiments have recently provided evidence that speech imagery is created by using forward models to generate a prediction of the sound of one's own voice (Tian & Poeppel, 2010, 2012; Scott, 2013a). This would suggest the possibility that inner speech can, as can visual and lexical information, induce recalibration. The experiments reported here test this possibility.

While the forward model account of recalibration is one of the theories that motivated the experiments reported here, these experiments are not a test of this theory; they are intended solely as a *proof of effect* that speech imagery can induce recalibration. These experiments do not distinguish between the various theories concerning the cause(s) of recalibration.

### Common source

One possible reason that recalibration would *not* be predicted to occur when the disambiguating source of information is speech imagery is that imagery introduces a competing potential 'source' for the perceived speech event.

In studies using visual information to induce recalibration, the video and audio are plausibly interpreted by perceivers as informative of the same speech event. In studies on lexical recalibration, there is only one sensory source and so a discrepancy between sources is not immediately relevant. However, in the illusion reported in Scott et al. (2013), where participants alter perception through imagery, the source of imagery is obviously different from that of the external sound. This discrepancy may be enough to prevent recalibration.[4]

This is related to the findings of Kraljic et al. (2008b). In that experiment, people did not show recalibration of an ambiguous /s/∼/ʃ/ sound if the person pronouncing the ambiguous sound had a pen in their mouth at the time and thus the unusual pronunciation could be attributed to that perturbation. Similarly, Kraljic et al. (2008a) found that recalibration occurred if people attributed the 'atypical' sound to an idiolectal variation, but not if there was a phonological environment that could be responsible for the atypicality of the sound. Thus, recalibration is quite subtle and seems to take into account the appropriateness of the inducer.

A similar issue of source sensitivity occurs with respect to visual recalibration. When a person is familiar with a face and voice, and thus can detect a mismatch between them, visual influences on speech perception are less likely to occur, again indicating that the degree to which the auditory information is perceived as belonging to the same source as the visual information affects whether the auditory perception is altered.

Thus the plausibility of a common source event for the auditory information and the context information (visual, lexical, reading imagery) may be a factor in whether recalibration occurs. In previous recalibration studies, there was no obvious mismatch between the source of auditory and context information, however that is less clearly the case here where the context is supplied by the participants' own imagery. The current experiment examines whether, despite this possible source mismatch issue, imagery can induce recalibration.

## Experiment one

The first experiment was designed to establish that recalibration can indeed be induced by speech imagery. For this experiment, only *enacted* speech imagery (silent "mouthing") was used. Participants were asked to mouth either /ˈibih/ or /ˈidih/ in synchrony with a sound that was ambiguous between these endpoints. Scott et al. (2013) demonstrated that this will result in the ambiguous sound being heard in line with what is being mouthed. After multiple exposures to this effect, if recalibration occurs, participants should continue to hear the ambiguous sound as matching what they had been mouthing even when they have ceased to mouth. Thus there were two components to the experiment, an exposure phase in which participants mouthed along with the ambiguous sound, and so heard that sound as what they were mouthing; then a test phase in which they heard the ambiguous sound on its own (without mouthing) and categorized it.

To ensure that any recalibration was genuinely a matter of remapping the category of an ambiguous sound (as opposed

[4] Of course, if all forms of recalibration are mediated by forward models, then the immediate source of information is always the perceiver's own motor system.

to a response bias or categorization 'inertia' on the part of participants), the experiment included two levels of stimulus clarity: **Ambiguous** and **Clear**. For the **Ambiguous** condition, participants were asked to mouth along with an ambiguous sound; however in the **Clear** condition, participants mouthed along with **un**ambiguous tokens of /ˈibih/ and /ˈidih/. In this condition, participants mouthed the *same* sound that they were hearing, so there was no mismatch between mouthing and exposure sound. Repeated exposure to an **un**ambiguous token of a sound is known to induce the converse effect of recalibration, *adaptation*, as discussed in "Recalibration".

In summary, one condition of the current experiment repeatedly exposes participants to an **Ambiguous** sound which should induce *recalibration*, while the other condition repeatedly exposes participants to a **Clear** sound which should induce *adaptation*. Finding both effects would show that recalibration can genuinely be induced by speech imagery and that the effect is not simply a matter of response bias or *categorization inertia*[5] since the actions of participants are the same in both conditions, only the clarity of the accompanying sound varies across conditions, and so any difference between conditions is attributable solely to the clarity of the exposure sounds.

## Methods

The experiment consisted of four conditions (2 levels of clarity, each with 2 sounds): **Clear** /ˈibih/,**Clear** /ˈidih/, **Ambiguous** /ˈibih/, **Ambiguous** /ˈidih/. These conditions were interleaved with each other (8 interleaved repetitions of each condition) – with the order counterbalanced across participants. With 4 conditions and 8 repetitions per condition, that leads to 32 blocks in this experiment. Half of the participants started with mouthing /ˈibih/ and the other half with mouthing /ˈidih/. Half of the participants started with **Clear** exposure sounds, half started with **Ambiguous** exposure sounds. The side of the response buttons was also counterbalanced across participants, leading to 8 versions of the experiment.

### *Stimuli*

A 10 000 step continuum between the (phonotactically possible) nonsense Arabic words /ˈibih/ and /ˈidih/ was created with STRAIGHT (Kawahara et al., 1999), using recordings of a Native Emirati Arabic speaker as the base.[6] STRAIGHT allows for a very natural-sounding continuum to be synthesized from real speech by segregating the pitch, intensity, duration, and spectrum of two natural sounds. These factors can then be specified at any intermediate compromise between the two sounds. In this experiment, the pitch, intensity and duration were set at a 50/50 compromise between the two endpoints and only the spectrum was adjusted from 99.97 % /ˈibih/ at the start of the continuum to 0.03 % /ˈidih/ at the end of the continuum.[7] The sounds were then filtered at 6800 Hz with a 100Hz roll-off (using Praat — Boersma and Weenink 2001) to eliminate some minor high-frequency distortion that was introduced by the synthesis process. The sounds were 478 ms long. The clear endpoints were used as the exposure sounds in the **Clear** condition.

Participants underwent a staircase procedure (Cornsweet, 1962) at the beginning of the experiment to estimate their 50 % point along this continuum which was used as the maximally ambiguous token for the **Ambiguous** condition. The staircase procedure consisted of two interleaved staircases with random switching, one starting at point 2400 on the continuum, the other at point 7600. There were 12 reversals with decreasing step size on each reversal except the last. The step sizes were: 1250, 700, 400, 250, 100, 50, 25, 10, 5, 2, 1, 1. The estimated perceptual boundary for each participant was calculated using a logistic regression over all of the data from their two staircases.

### *Participants*

There were 24 female participants (average age = 20.1 years; SD = 1.26 years), all were paid or given course credit for their participation. All participants were students of United Arab Emirates University and were native speakers of Emirati Arabic. One participant's boundary was too erratic to be fit by logistic regression. She was not run in the main experiment and was replaced to reach a total of 24 participants.

[5]The tendency for people to continue to categorize ambiguous stimuli in line with recent experience — see the **General Discussion** for further explanation.

[6]While a 10 000 step continuum may seem excessive, it should be kept in mind that there is no 'cost' to having such a finely divided continuum available, and it allows for more precise matching of tokens to participants' estimated phoneme boundaries. Each participant, of course, only heard a fraction of these steps, so the number of continuum steps is irrelevant to the structure and duration of the experiment.

[7]These are the closest values to 0 % and 100 % that the software allows.

*Procedure*

In each block, there was an exposure phase in which participants were exposed to a sound ambiguous between /ˈibih/ and /ˈidih/(the **Ambiguous** condition); or which was a clear instance of /ˈibih/ or /ˈidih/ (the **Clear** condition). The ambiguous sound corresponded to each participant's estimated boundary between /ˈibih/ and /ˈidih/ as determined by the staircase procedure. There were 26 such exposures in each block, presented with 465ms ISI. Participants silently mouthed /ˈibih/ or /ˈidih/in time to these sounds.

As participants had to time their mouthing to be in synchrony with the ambiguous sound during the exposure phase, a visual aid to the rhythm of sound presentation was used. A pulsing red circle (similar to a karaoke ball) appeared on screen during the exposure phase, pulsing in time with the presented sounds (growing bigger or smaller in synchrony with the amplitude of the sound). As most people are familiar with "lip-synching" to a rhythmic sound, it is unlikely that keeping this rhythm was a difficult task.

The instructions to participants were to "lip synch" to the sound just as they would if they were mouthing along with a familiar song, and that they should "hear their voice in their head" while doing this.

Following the exposure phase, there was a 1.35 second pause after which participants categorized 6 audio-only tokens of ambiguous /ˈibih/∼/ˈidih/ sounds. The sounds presented in this test phase corresponded to each participant's estimated 50 % point in perceptual space between /ˈibih/ and /ˈidih/. The same maximally ambiguous sound was used throughout the test phase to maximize the perceptual ambiguity and thus increase power.

Each condition was presented to the participants 8 times throughout the experiment (in interleaved blocks), with 6 targets per block, resulting in 48 target categorizations per condition.

Each participant was run separately in a quiet room of the *Perceptual Laboratory* at United Arab Emirates University. Stimuli were presented over headphones. The experiment was run on the software *Psychopy* (Peirce, 2009). The main experiment was preceded by an identically structured practice session (lasting c. 3 minutes) to familiarize participants with the procedure.

In order to help ensure participants were paying attention to the task, a 2-letter code had to be typed in by the participant to start each block. The code consisted of one letter for the action to be performed ('m' for "mouthing") and one letter for the sound to be mouthed ('b' or 'd' for /ˈibih/or /ˈidih/). The text informing participants of the sound to be mouthed was not displayed during the exposure phase and so recalibration from text, as recently demonstrated by Keetels et al. (2016), is not plausible in this experiment (or any of the experiments reported in this paper). During the experiment a research assistant did a regular check of whether participants were following instructions carefully – silently mouthing the appropriate sounds, in rhythm, during the exposure phase and not mouthing during the test phase.

## Results

A 2 X 2 repeated-measures ANOVA using the 'ez' package (Lawrence, 2013) in R (R Core Team, 2014) was performed with "Clarity" factors: **Clear** and **Ambiguous**, and "Sound" levels: /ˈ**ibih**/and /ˈ**idih**/. There was a significant interaction of Clarity and Sound [$F(1,23) = 28.493$, $p < 0.001$]. Planned t-tests showed that more/ˈidih/ were perceived after repeated exposure to /ˈidih/ in the **Ambiguous** condition ($p = 0.0027$), but the converse effect (though only marginally significant), more /ˈidih/ perceived after repeated exposure to /ˈibih/, in the **Clear** condition ($p = 0.05045$) (Fig. 1).

## Discussion

The results clearly show that in the test phase participants heard *more* of the sound they had been mouthing during the exposure phase when the accompanying exposure sound was ambiguous, but when the accompanying exposure sound was clear, participants heard *fewer* instances of the sound they had been mouthing. This shows that mouthing causes recalibration for ambiguous sounds but that clear sounds induce adaptation, exactly as predicted.

It should be noted that the adaptation in the **Clear** condition was only marginally significant. This limits the claims that can be made about adaptation for this data set. However, the interaction between Clarity and Sound was significant, indicating that people perceive the target sounds differently depending on the clarity of the sounds in the exposure phase.

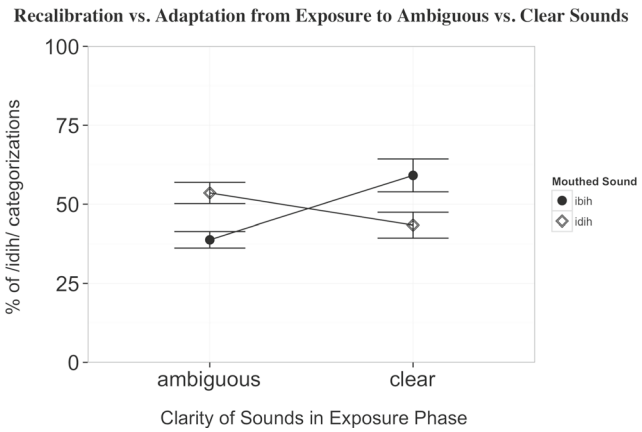


**Fig. 1** Results of Experiment One — SE bars are shown

This is strong evidence that the recalibration is not simply a matter of response bias or *categorization inertia*, as participants' actions were identical in the exposure phases of both conditions, as were the categories perceived during the exposure phases. Despite this equivalence, their responses differed in the test phases of these conditions indicating that the difference in their responses was *not* determined by what they were doing in the exposure phase *nor* by the category they perceived during the exposure phase. Indeed, participants actions changed between exposure and test phases (they mouthed during the exposure phase, but **not** during the test phase) thus there was a clear change in action between exposure and test phases which should have emphasized that the situation had altered and so help interfere with any behavioural 'inertia'.

There remains the possibility that the imagery in the **Ambiguous** condition is necessarily more effortful when the simultaneous speech sound is ambiguous (and so less in agreement with participants' imagery) than when imagery and speech sound are in agreement (in the **Clear** condition) and that this increased effort leads to a larger response-bias.

While such an interpretation of the current experiment is possible, it requires two assumptions: that speech imagery is more effortful when accompanied by ambiguous sound and that such greater effort would lead to an increased response-bias. Inasmuch as the same situation holds for visually or lexcially-induced recalibration, this alternative is not just a different interpretation of the current experiment but of recalibration in general.

The claim that categorization is more effortful in the ambiguous condition is quite plausibly true, however the same interpretation holds for demonstrations of recalibration from visually induced recalibration — It has been shown that response-times to audiovisual stimuli with ambiguous audio are slower than to audiovisual stimuli with unambiguous audio (Massaro et al., 1993). Similarly, the influence of lexical information on perception (the "Ganong" effect) is also associated with slower response times (Fox, 1984; Pitt, 1995). This suggests that categorization is more effortful whenever the auditory signal is ambiguous. Thus, if the current results may be interpreted as categorization inertia, then the same interpretation could be applied to the vast literature demonstrating recalibration from visual and lexical information as well.

This issue is in some sense a matter of interpretation, however, as recalibration may be considered (no matter what its inducer) a form of categorization inertia in that its defining characteristic is a tendency to continue to categorize sounds in line with how one had categorized them in recent experience. The current set of experiments simply demonstrate that speech imagery can induce a form of recalibration that appears qualitatively indistinguishable from the currently known inducers. This issue will be taken up again in **Experiment Two**, where imagery-induced and visually-induced recalibration are shown in the same experiment.

This experiment thus establishes that speech imagery can induce recalibration. This is replicated and extended in **Experiments Two** and **Three**.

## Experiment two

Scott et al. (2013) showed that both mouthed and pure inner speech can alter concurrent speech perception. **Experiment One** demonstrated recalibration from mouthed speech imagery. The current experiment follows up this finding by examining whether pure speech imagery can also have this effect. Furthermore this experiment compares the recalibration caused by both forms of speech imagery with that caused by visual information.

### Methods

The structure was similar to that of **Experiment One**, however as recalibration from imagery was established in **Experiment One**, there was no need to replicate the **Clear** control condition, so only ambiguous sounds were used in the exposure phase in **Experiment Two** — this was also necessitated by duration considerations.

Three methods of recalibration were tested in this experiment: "mouthed" speech imagery (**Mouth**), "pure" speech imagery (**Imagine**) and visual information (**Watch**); each tested with the two sounds /ˈibih/ and /ˈidih/ giving a 3 X 2 structure of conditions.

These conditions were interleaved, with the order counterbalanced across participants (8 interleaved repetitions of each condition throughout the experiment). All 6 possible order permutations of the Action factor were used — for each of these, half of participants started with /ˈibih/ and the other half with /ˈidih/. The side of the response buttons was also counterbalanced across participants, leading to 24 versions of the experiment. There were 48 blocks in total (6 conditions with 8 repetitions per condition).

#### *Stimuli*

The same 10 000 step continuum was used as in **Experiment One**. The same staircase procedure was again used to estimate participants' 50 % point in perceptual space between /ˈibih/ and /ˈidih/. The same native speaker of Emirati Arabic who provided the audio recordings was video-recorded saying /ˈibih/ and /ˈidih/ for use in the **Watch** condition in this experiment.

*Participants*

There were 24 new female participants (average age = 22.1 years; SD = 3.47 years). All were paid or given course credit for their participation. All participants were students of United Arab Emirates University and were native speakers of Emirati Arabic. One participant failed to follow instructions (as determined by the research assistant administering the experiment) during the test phase and so the research assistant halted the experiment early. This participant was thanked and paid, but her data set was deleted without being examined. She was replaced to ensure a total of 24 participants.

*Procedure*

In each block, there was an exposure phase in which participants were exposed to a sound ambiguous between /ˈibih/and /ˈidih/. This sound corresponded to each participant's estimated boundary between /ˈibih/ and /ˈidih/. There were 10 exposures in each block presented with 465 ms ISI.[8]

On each exposure, the ambiguous sound was disambiguated to one of these categories (always to the same category during a given block) because of the accompanying *mouthing* (**Mouth**), *pure imagery* (**Imagine**), or *video* (**Watch**).

During the **Mouth** and **Imagine** conditions, the same visual aid as in **Experiment One** (a pulsing red circle) was presented on the screen in synchrony with the ambiguous sound in the exposure phase. During the exposure phase of the **Watch** condition, the pulsing circle was replaced by a video of a face pronouncing the sound /ˈibih/ or /ˈidih/.

The instructions to participants about mouthing were as in **Experiment One**. The instructions on imagining were to remind them of how we all "talk to ourselves in our head" and that the goal was for the participant to hear her own voice in her head saying the appropriate sound in time to the external sound. They were reminded that this was just like the **Mouth** condition, but "without actually moving your mouth". For the **Watch** condition, they were asked to neither imagine, nor mouth, just to watch the video.

Following the exposure phase there was a 1.35 second pause after which participants categorized 6 audio-only tokens of ambiguous /ˈibih/∼/ˈidih/ sounds. The sounds presented in the test phase corresponded to each participant's 50 % point in perceptual space between /ˈibih/ and /ˈidih/.

Each condition was presented to the participants 8 times throughout the experiment (in interleaved blocks), with 6 targets per block, resulting in 48 target categorizations per condition. The main experiment was preceded by an identically structured practice session (lasting c. 3-4 minutes) to familiarize participants with the procedure.

In order to ensure participants were paying attention to the task, a 2-letter code had to be typed in by the participant to start each block. The code consisted of one letter for the action to be performed ('m' for "mouthing", 'w' for "watching" or 'i' for "imagining") and one letter for the sound to be mouthed ('b' or 'd' for /ˈibih/ or /ˈidih/. As with **Experiment One** a research assistant regularly checked whether participants were following instructions.

## Results

A 3 X 2 repeated-measures ANOVA (using the 'ez' package in R) was performed with "Action" factors: **Mouth**, **Imagine** and **Watch**, and "Sound" levels: /ˈibih/ and /ˈidih/. There was a significant main effect of Sound. There was also a significant interaction of Action and Sound [$F(2, 46) = 29.753$, $p < 0.001$]. Planned t-tests showed that more /ˈibih/ were perceived after repeated exposure to /ˈibih/ in the **Watch** condition ($p < 0.001$), **Mouth** condition ($p=0.002$) and **Imagine** condition ($p < 0.001$) (Fig. 2).

As all three Action factors showed a significant difference between /ˈibih/ and /ˈidih/ in the same direction, the interaction between Sound and Action is a matter of strength of the recalibration. A follow-up multilevel model analysis (linear mixed-effects model fit by maximum likelihood) was conducted using the 'nlme' package in R (Pinheiro

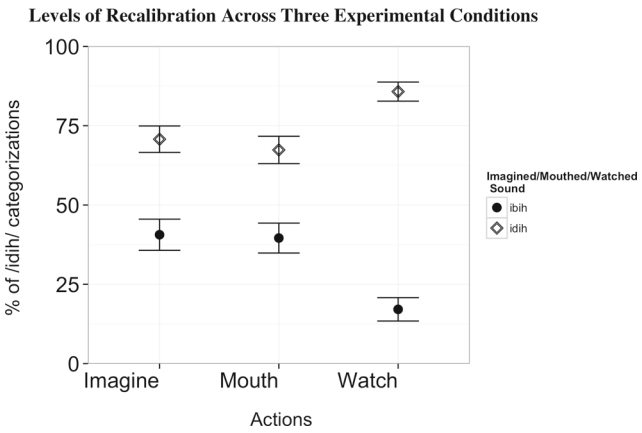


**Fig. 2** Results of Experiment Two — SE bars are shown

[8]This experiment involved 50 % more conditions than **Experiment One** and so to keep the duration of the experiment within manageable limits, the number of exposure tokens was reduced from **Experiment One**. In addition, Vroomen et al. (2007) showed that above 8 exposures there is little increased recalibration effect.

et al., 2014). This analysis was conducted to look at the difference in degree of recalibration between the **Watch**, **Mouth** and **Imagine** conditions (i.e. the interaction between Sound and Action). This analysis replicated the significant main effect of Sound [$\chi^2(1) = 44.103$, $p < 0.001$]; and the significant interaction of Sound by Action [$\chi^2(2) = 57.211$, $p < 0.001$].

Orthogonal contrasts revealed that the shift in /ˈibih/-responses between the /ˈibih/ vs. /ˈidih/ Sound factors (i.e. the recalibration) was significantly larger in the **Watch** condition compared to the other two conditions (**Mouth** and **Imagine**) [$b = 6.619$, $t(92) = 8.646$, $p < 0.001$]. There was no significant difference in the strength of recalibration in the **Imagine** vs. **Mouth** conditions [$b = -0.5859$, $t(92) = -0.4419$, $p = 0.659$].

As discussed in **Experiment One**, one possible interpretation of the imagery-induced recalibration results is that imagery is more effortful when the accompanying audio is ambiguous and thus may be more likely to cause a response-bias in the form of *categorization inertia* in which people continue to categorize sounds in line with previous categorizations, even when something in the situation has changed. This experiment directly compares visual and imagery inducers of recalibration. If these inducers constitute qualitatively different phenomena, we might expect that the duration of the effect would be different between visual and imagery conditions (Arthur Samuel has suggested, in personal communication, that categorization inertia should show a steeper decline in effect over the course of response trials than recalibration). To examine whether the decline of recalibration over the course of response trials was different for visual and imagery conditions, the level of recalibration for each of the 6 response tokens was plotted and Pearson's product-moment correlation was calculated separately for each condition. The results are shown in Figs. 3, 4, and 5.

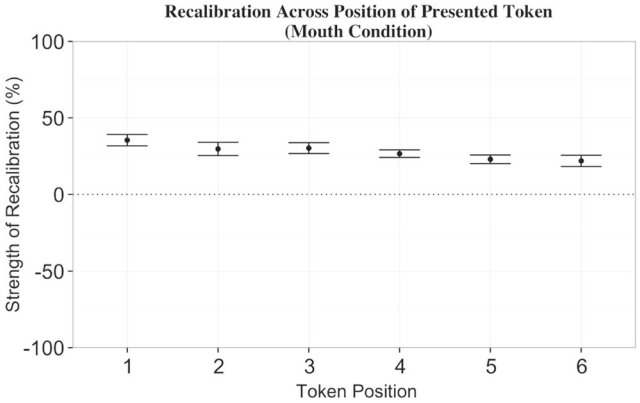


**Fig. 4** Experiment 2 Recalibration by Token (**Mouth** Condition) — SE bars are shown

Spearman's rank correlation was calculated for **Strength of Recalibration** by **Token Position** in each of the Action conditions: [**Watch** condition ($rho = -0.0774$, $p = 0.356$), **Mouth** condition ($rho = -0.1156$, $p = 0.1678$), **Imagine** condition ($rho = 0.0076$, $p = 0.9282$)].

As can be seen in the graphs and correlation analyses, the **Watch** and **Mouth** conditions show a small, non-significant, decline in recalibration across the test phase and the **Imagine** condition shows an even smaller (and non-significant) *increase* in recalibration. This gives no reason to think that the recalibration in the **Mouth** and **Imagine** conditions is of a qualitatively different kind from that in the **Watch** condition. In all three conditions there was a non-significant change across the 6 positions of the test-phase. This supports the argument, presented in "Discussion", that the recalibration demonstrated for imagery in this experiment is qualitatively the same type of phenomenon as the well-established visually induced recalibration.

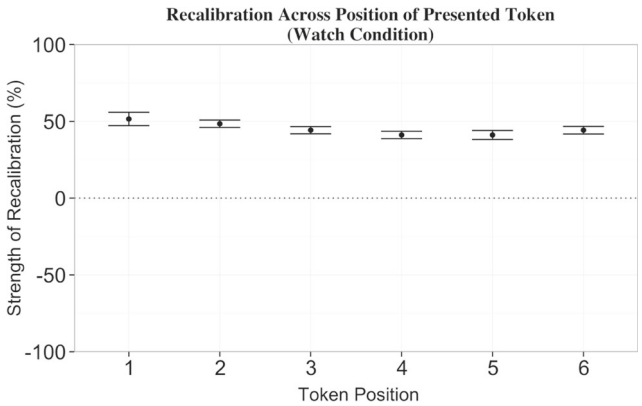


**Fig. 3** Experiment 2 Recalibration by Token (**Watch** Condition) — SE bars are shown

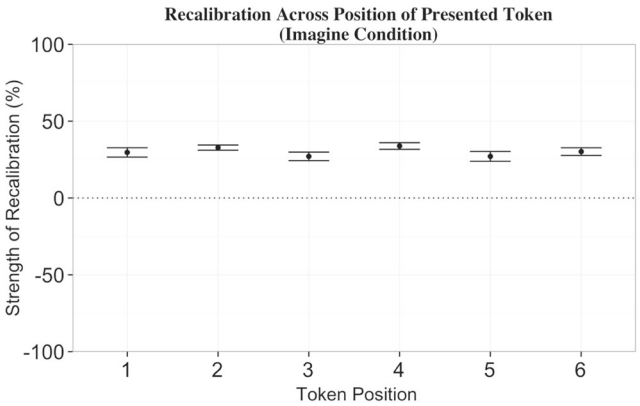


**Fig. 5** Experiment 2 Recalibration by Token (**Imagine** Condition) — SE bars are shown

### Discussion

This experiment extends the findings of **Experiment One** to the case of "pure" speech imagery (without any movement of the speech articulators), showing that pure inner speech can also induce recalibration.

The levels of recalibration from pure and mouthed inner speech were comparable in this experiment but were significantly lower than that induced by visual information. This could indicate a different mechanism or it could simply be an indication of the weaker impact of speech imagery on perception. In the original experiment showing an influence of speech imagery on perception, Scott et al. (2013) found that imagery, in the strongest condition, only caused a shift c. 80 % of the time which, while significant, does not reach the reliability of visual influence on speech perception, which is famously robust. Alternatively, the weaker impact of both forms of imagery may be because of the fact that, as discussed in "Common source", imagery is clearly originating from a different source from the ambiguous sounds in the exposure phase (unlike the video in the **Watch** condition).

In the original Scott et al. study, mouthing produced a significantly larger effect than pure imagery, and so we might expect to see a larger degree of recalibration from mouthing than from pure imagery. In this experiment (and in **Experiment Four**) this is not the case. One tentative explanation for this is that the greater degree of phonetic detail present in mouthed speech imagery (Oppenheim and Dell, 2010) interferes with the remapping of sound to category. Specifically, recalibration involves learning a mapping between certain phonetic details and a particular (phonological) category. If those phonetic details are obscured (or "overwritten") to some degree by the phonetic content of the illusion that is inducing the recalibration, then the recalibration would likely be weaker than would otherwise be expected. This fits with the current understanding of mouthed vs. pure speech imagery which finds that pure imagery is largely devoid of phonetic content whereas mouthed imagery contains such content (Oppenheim & Dell, 2010). It should be noted that this explanation is highly tentative and requires further research to establish its (in)correctness.[9]

[9]Though I would like to point out that the possibility of a weaker recalibration from mouthed speech for this reason had been considered *before* these experiments were run. As the purpose of these experiments is to provide an initial demonstration of recalibration from speech imagery, the exploration of this issue is left to future research.

## Experiment Three

As discussed in "Recalibration", recalibration seems to be an issue of mapping an ambiguous sound to a particular category and recalibration will occur whenever the perceiver has some way (lexical information, visual information or, as in the experiments reported here, imagery information) to know the category to which the ambiguous sound should be linked. If it is the knowledge of the category that is relevant, and not the particular pathway to that knowledge, then we should expect that simply informing perceivers of the identity of the sound should induce recalibration. **Experiment Three** tests this.

In this experiment, participants were merely told that they would hear several repetitions of either /ˈibih/ or /ˈidih/ and that they would then be asked to categorize some ambiguous sounds (as either /ˈibih/ or /ˈidih/). The prediction is that this should induce recalibration, though, as simply being told the category of a sound is presumably a less reliable way to alter perception than being "tricked" by a perceptual illusion, the strength of this recalibration is predicted to be weaker than recalibration from imagery.

### Methods

The Stimuli and procedures for this experiment were identical to those in **Experiment Two** with the exception that participants were simply told that they would hear either /ˈibih/ or /ˈidih/ repeated several times. No mention was made of mouthing or imagining. Thus there were only two conditions in this experiment: **Believe /ˈibih/** (Participants being told the ambiguous sound was /ˈibih/) and **Believe /ˈidih/** (Participants being told the ambiguous sound was /ˈidih/).

All other aspects of the experiment were identical with **Experiment Two**. The participants were reminded (by onscreen text) at the start of every exposure block which sound they would be hearing in that block. A two-letter code was again used. This time the first letter was "l" for "listen" followed by the letter for the sound to be heard ('b' or 'd' for /ˈibih/ or /ˈidih/).

#### *Participants*

There were 24 new female participants (average age = 20.82 years; SD = 1.63 years). All were paid or given course credit for their participation. All were students of United Arab Emirates University and were native speakers of Emirati Arabic. Three participants' boundaries were too erratic to be fit by logistic regression and so were not run in the

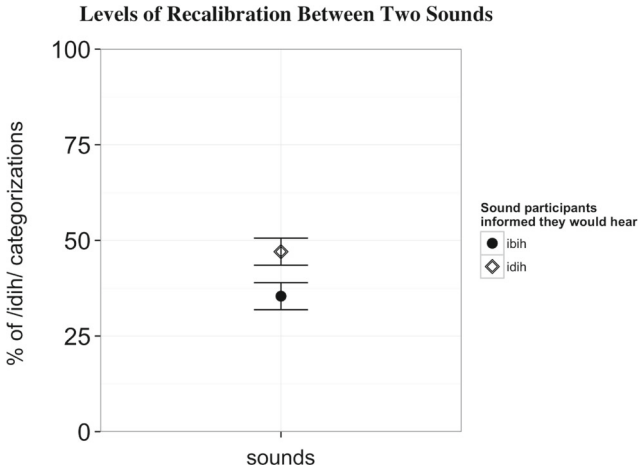


**Fig. 6** Results of Experiment Three — SE bars are shown

main experiment and were replaced to reach a total of 24 participants.

## Results

A paired t-test was conducted between the **Believe** /ˈibih/ and **Believe** /ˈidih/ conditions. This was significant [$t = -2.3263$, $df = 23$, $p = 0.02917$] (Fig. 6).

As can be seen, simply being told the category of the ambiguous sound during the exposure phase was sufficient to induce a small degree of recalibration. This supports the contention that recalibration relies on mapping ambiguous sounds to categories, and as long as the target category can be determined, recalibration can occur.

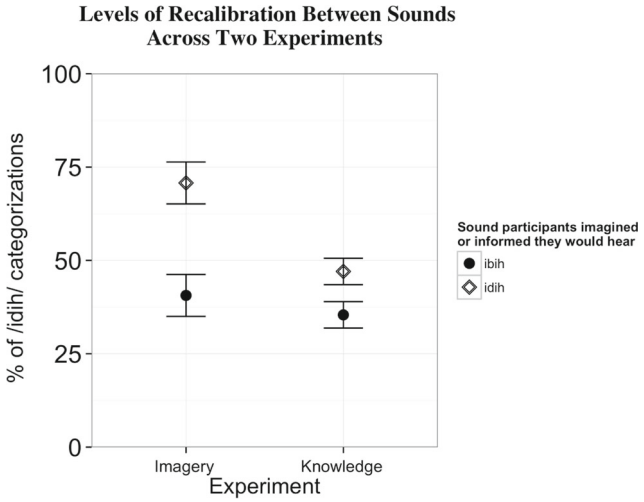


**Fig. 7** Results of Comparison between Experiment Two (Imagery Condition) and Experiment Three — SE bars are shown

It would be expected that perceptual illusions are a more reliable method of causing people to hear an ambiguous sound as a particular category and so should be a more reliable way to induce recalibration. In order to compare the strength of recalibration in this experiment with that seen in the **Imagery** condition of **Experiment Two**, a mixed-model ANOVA (using the "ez" package in R) was run with the within-subjects factor **Sound** and the between-subjects factor **Experiment**. This found a significant effect of **Experiment** [$F(1, 46)=13.979418$, $p < 0.0001$] (Fig. 7). This shows that while simple knowledge was sufficient to induce recalibration, the strength of this recalibration was significantly weaker than recalibration induced by imagery.

## Discussion

Recalibration from being given explicit instruction as to the identity of the ambiguous sound was found in **Experiment Three**. This is not a surprising result, as it has been argued (Reinisch et al., 2014) that recalibration is a matter of manipulating the perceiver's knowledge of the category of the sound (no matter how that knowledge is achieved). Indeed, previous research has already established that being told the identity of a sound does have effects on auditory perception. A compelling example of this is the "White Christmas" effect (Merckelbach & Ven, 2001) in which some people can be induced to hear the song "White Christmas" when presented with white noise, simply by telling them that the song might be buried under the noise (the song is, in fact, not present). The current experiment uses essentially the same methodology to show that this explicit knowledge can have an after-effect on perception in the form of recalibration. The strength of this recalibration from explicit knowledge, though, is not as strong as from the perceptual effects of speech imagery.

This suggests that the key to recalibration is knowledge of the target category to which the ambiguous sound should be associated. It could be argued that the recalibration shown in this experiment should be considered a special case of lexical recalibration in that both rely on the knowledge of the perceiver. As with lexical knowledge, there remains the possibility (not tested in these experiments) that forward models are acting as the mediating mechanism.

While it is clear that knowledge of the target category is central to recalibration, it is also clear that not all routes to this knowledge are equally effective. Imagery induces a significantly larger degree of recalibration than simply being told the category, as shown in this experiment. Furthermore, visual information has a larger impact on recalibration than does imagery, as shown in **Experiment Two** and

**Experiment Four**. Future research will have to address the reasons for these differing degrees of impact.

## Experiment four

**Experiment Four** is an extension of **Experiment Two**, replicating the recalibration found there, but using a new set of (fricative) speech sounds. This demonstrates that the recalibration found in **Experiments One** and **Two** extends to a new class of sounds.

The replication of **Experiment Two** with fricatives is not a given. The articulation of fricatives is quite different from stops in ways that may interact with imagery (mouthed or pure). Specifically, fricatives are more dependent on airflow than are stops — the exact position of the articulators and the narrowness of the constriction for fricatives relies on the presence of airflow during normal speech in a way that is not true of stops (Stevens, 2000). As this airflow is absent in speech imagery, imagery of fricatives may be impaired in comparison to imagery of stops. Similarly, fricatives are more dependent on precise sensorimotor control than are stops (Borden et al., 1973; Ladefoged & Maddieson, 1996). In pure imagined speech, with no overt articulation, there is no somatosensory feedback. The absence of this feedback may interfere with imagery for fricatives. Thus, fricatives are different from stops in ways that are relevant to speech imagery and so it is important to show that the recalibration effect demonstrated for stops in **Experiments One**, **Two** and **Three** is also present for fricatives.

### Methods

The Methods were identical to those in **Experiment Two** with the exception of the speech sounds involved, which, for this experiment, were the (phonotactically possible) nonsense Arabic words /ˈifih/ and /ˈiθih/.

#### *Stimuli*

A 10 000 step continuum between the nonsense Arabic words /ˈifih/ and /ˈiθih/ was created in the same way and using the same speaker, as in **Experiments One** and **Two**. The speaker was also video recorded for the stimuli in the **Watch** condition. At the beginning of the experiment, participants underwent the same staircase procedure as in **Experiments One** and **Two** to estimate their 50 % point along the continuum. The sounds were 635ms long.

#### *Participants*

There were 24 new female participants (average age = 20.1 years; SD = 1.66 years), all were paid or given course credit for their participation. All participants were students of United Arab Emirates University and were native speakers of Emirati Arabic. Five participants failed to follow instructions (as determined by the research assistant administering the experiment) during the test phase and so in each case the research assistant halted the experiment early. These participants were thanked and paid, but their data sets were deleted without being examined. They were replaced by new participants to ensure a total of 24 participants.

### Results

A 3 X 2 repeated-measures ANOVA, using the 'ez' package in R, was performed with "Action" factors: **Watch**, **Mouth**, **Imagine** and "Sound" levels: /ˈifih/ and /ˈiθih/. There was a significant main effect of Sound [$F(1, 23) = 92.7979$, $p < 0.001$]. There was a significant interaction of Action and Sound [$F(2, 46) = 29.753$, $p < 0.001$]. Planned t-tests showed that more /ˈifih/ were perceived after repeated exposure to /ˈifih/ in the **Watch** condition ($p < 0.001$), **Mouth** condition ($p=0.002$) and **Imagine** condition ($p <0.001$).

As all three Action factors showed a significant difference between /ˈifih/ and /ˈiθih/ in the same direction, the interaction between Sound and Action is a matter of strength of the recalibration. A follow-up multilevel model analysis (linear mixed-effects model fit by maximum likelihood) was conducted using the 'nlme' package in R. This analysis was conducted to look at the difference in degree of recalibration between the **Watch**, **Mouth** and **Imagine** conditions (i.e. the interaction between Sound and Action).

This analysis replicated the significant main effect of Sound [$\chi^2(1) = 71.155$, $p < 0.001$]; and the significant interaction of Sound by Action [$\chi^2(2) = 13.42270$, $p = 0.0012$] (Fig. 8).

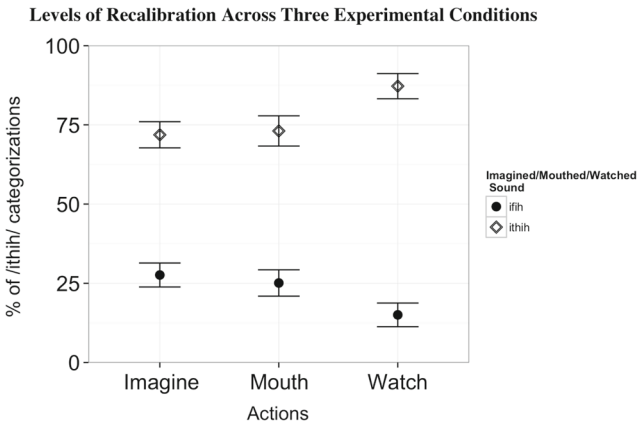


**Fig. 8** Results of Experiment Four — SE bars are shown

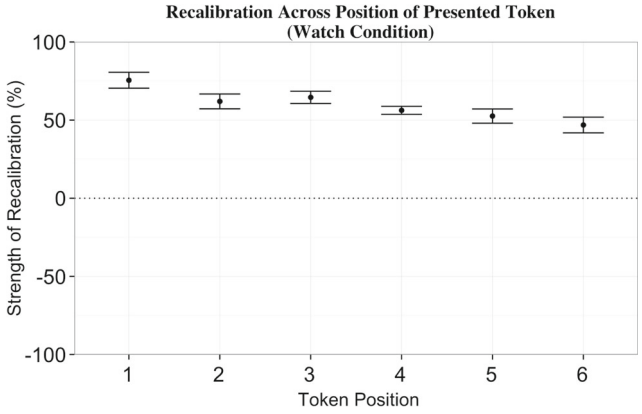


**Fig. 9** Experiment 4 Recalibration by Token (**Watch** Condition) — SE bars are shown

Orthogonal contrasts revealed that the shift in /ˈifih/- responses between the /ˈifih/ vs. /ˈiθih/ Sound factors (i.e. the recalibration) was significantly larger in the **Watch** condition compared to the other two conditions (**Mouth** and **Imagine**) [b = 0.5715, t(92) = 3.6935, p < 0.001]. There was no significant difference in the strength of recalibration in the **Imagine** vs. **Mouth** conditions [b = −0.1085, t(92) = −0.4049, p = 0.6865].

As with **Experiment Two**, Spearman's rank correlation was calculated for **Strength of Recalibration** by **Token Position** in each of the Action conditions in order to examine whether there was a significant decline in recalibration across the 6 tokens of the test-phase. This was to check for the possibility that *categorization inertia* can explain these results because of the possibility that *categorization inertia* would show a sharper decline than recalibration: [**Watch** condition (rho = −0.1908, p= 0.0219), **Mouth** condition (rho = −0.1147, p = 0.1711), **Imagine** condition (rho= −0.1724 , p = 0.0388 )] (Figs. 9, 10 and 11).

As in **Experiment Two** the decline in recalibration by token position is small, though in two of the conditions (**Watch** and **Imagine**) it did reach significance in this experiment (unlike in **Experiment Two**). However, it should be noted that the decline in the **Watch** condition is marginally *larger* than in the **Imagine** condition (and larger than the non-significant **Mouth** condition). This contradicts the argument that recalibration from imagery (mouthed and pure), but *not* from visual information, is due to *categorization inertia*, which would predict that decline in recalibration would be sharper in the imagery conditions than in the visual condition. Thus, this experiment, as with **Experiment Two**, provides no evidence that the recalibration is qualitatively different when it is caused by visual information versus imagery. This issue is taken up again in the **General Discussion**.

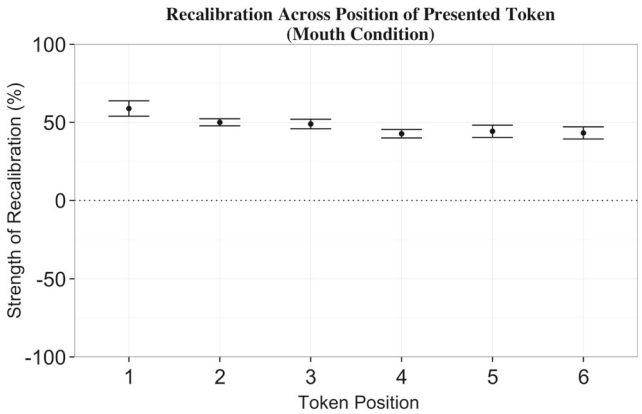


**Fig. 10** Experiment 4 Recalibration by Token (**Mouth** Condition) — SE bars are shown

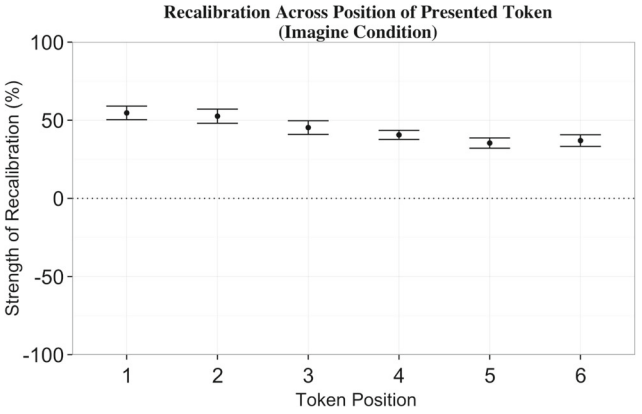


**Fig. 11** Experiment 4 Recalibration by Token (**Imagine** Condition) — SE bars are shown

### Discussion

This experiment replicates the recalibration shown in **Experiments One** and **Two** and extends the finding to fricatives. This shows that recalibration from speech imagery is not exclusive to stops. The extension of recalibration to fricatives is an important finding as fricatives are more reliant than stops (in normal speech) on airflow and kinaesthetic feedback (Stevens, 2000; Borden et al., 1973; Ladefoged & Maddieson, 1996), and airflow is absent in imagined speech (both pure and mouthed) and kinaesthetic feedback is absent in pure speech. This experiment also replicated the finding of **Experiment Two** that the recalibration induced by speech imagery was comparable for mouthed and pure speech imagery but both showed a weaker effect in comparison to recalibration induced by visual information.

## General discussion

These experiments show that imagery, in both enacted and non-enacted forms, induces recalibration. Thus, the

perceptual effect demonstrated in Scott et al. (2013) does not just influence simultaneous perception but induces a perceptual shift that lingers after the imagery has stopped. In **Experiments One** and **Two** this recalibration was shown for imagery (both mouthed and pure) of stop consonants.

Fricatives are more dependent than stops on the presence of airflow and on kinaesthetic feedback during production (Borden et al., 1973; Ladefoged and Maddieson, 1996; Stevens, 2000). Thus it may be possible that when airflow and kinaesthesis are altered or absent (as in imagery), fricatives may be in some way impaired in comparison to stops. Despite this possible issue in the imagery of fricatives, recalibration from fricatives was clearly demonstrated in **Experiment Four**.

The inclusion of the **Watch** condition in **Experiments Two** and **Three** allowed for the impact of imagery (both pure and mouthed) to be compared with the impact of visual information. These results show that imagery is able to induce recalibration, but not as strongly as visual information.

**Experiment Three** demonstrated that merely telling participants to which category an ambiguous sound belongs is sufficient to induce recalibration. This is in line with an implicit assumption in the field (Reinisch et al., 2014) that recalibration is dependent on the mapping of ambiguous phonetic information onto a phonological category and that any source of information that can indicate the intended phonological category should be able to induce recalibration. While **Experiment Three** supports this claim, it is obvious that not all sources of information cause the same level of recalibration. The level of recalibration from simply telling people the category of the ambiguous sound was significantly smaller than from pure imagery; similarly, in **Experiments Two** and **Experiment Four** it was shown that recalibration from visual information is significantly stronger than recalibration from either form of imagery.

These results are important for several reasons. First, recalibration is a perceptual effect with a vast literature that has concentrated on two primary inducers: visual and lexical (and very recently reading has also been shown to induce recalibration). This experiment demonstrates that another inducer of recalibration exists – speech imagery.

Second, the influence of speech imagery on the perception of external speech is a relatively new discovery and its parameters and the time-course of its effects need to be determined. These experiments demonstrate that the impact of speech imagery on perception can linger after the imagery itself has ended, in the form of recalibration.

Finally, these experiments demonstrate that the possible issue of source discrepancy between sound and disambiguating information (discussed in "Common source") does not prevent imagery from causing recalibration. Kraljic et al. (2008b) showed that recalibration is sensitive to context, and may not occur if the source of the ambiguity of the sound can be attributed to something other than the speaker's normal production. In the case of speech-imagery investigated here, the imagery is being produced by the listener while listening to an external sound and so the listener is clearly aware that the external sound and the imagery originate from different sources. Despite this conflict of sources, speech imagery still induced recalibration. This is unlike demonstrations of visual influence on speech perception in which the video and audio are typically aligned so that they seem to indicate a common source. This suggests that the sensitivity of recalibration to issues of information source is quite subtle and may not be completely cognitively penetrable, as the participants in the experiments reported here presumably were well aware that the source of the sound they were hearing was distinct from their own imagery, yet despite this awareness, recalibration still occurred. However, this source discrepancy may be responsible for the significantly weaker recalibration in the imagery conditions (in comparison to the **Watch** condition) of **Experiments Two** and **Four**, as mentioned in "Discussion".

As there is significant evidence that speech imagery is generated by means of forward models (Tian & Poeppel, 2010, 2012; Scott, 2013a), these experiments are consistent with the theory proposed in "Forward models" that lexical, visual and imagery-induced recalibration all share a common mechanism: "filling-in" from forward models. However, these experiments are intended merely as a *proof of effect* that speech imagery can induce recalibration and do not directly test the possibility that these different forms of recalibration are mediated by a common mechanism.

This raises the issue of what might be called "categorization inertia" — the tendency to continue to categorize an ambiguous stimulus in line with previous categorizations. Such inertia is not a new phenomenon, Leeper (1937 – cited in Wilton 1985) demonstrated that if people were presented with an ambiguous line drawing, they tended to categorize it in line with a similar, less ambiguous, line drawing they had previously seen. Indicating that, in the absence of some reason to change their mind, people will assume that the similar figure should continue to receive the same categorization. Perhaps the recalibration found in these studies is due to such inertia. Note that this is not exclusively an issue for this set of experiments; many previous demonstrations of speech recalibration could be interpreted in this way. To some degree, this is a matter of debating nomenclature, as recalibration does involve an inertial tendency to carry on hearing a sound in line with recent experience. However the distinction that may be drawn is that categorization inertia may be considered a response bias, whereas recalibration is apparently a matter of remapping (in the short term) of phonetic boundaries. Assuming

that existing methods of demonstrating recalibration do indeed show such a remapping, then the evidence suggests that the recalibration induced by imagery in the experiments reported here is comparable to existing instances of recalibration.

First, there is the evidence from **Experiment One** in which recalibration was pitted against adaptation. Here, categorization inertia from speech imagery clearly did not induce a recalibration-like shift in perception. When the participants imagined a speech sound repeatedly in time with a clear instance of the speech sound, despite the response-bias towards continuing to hear that speech sound that this should induce, they in fact reported hearing *fewer* instances of the imagined speech sound during testing (adaptation). This suggests that response-bias from imagery is not at work here, at least when the underlying auditory stimulus is unambiguous. When the underlying speech sound was ambiguous, the predicted recalibration was found. One might argue that when speech imagery clearly matches the speech sound being heard, then less effort is required and so a response bias is less likely to occur. This is certainly a possibility, but is not an interpretation unique to recalibration from imagery. Categorizations are typically slower (and so arguably more effortful) when audio is ambiguous (Massaro et al., 1993; Fox, 1984; Pitt, 1995) and so much of the recalibration literature (which relies on such ambiguity) could be interpreted under this categorization-inertia explanation.

Second, there is the analysis of the duration of the recalibration effect. Arthur Samuel (personal communication) has argued that a response-bias should show a rapid decline over the course of the test trials. However, the small decline in recalibration over the 6 token positions in these experiments was not significant. More importantly, the decline in recalibration found for visually-induced recalibration (an uncontroversial and well-established method to induce recalibration) is indistinguishable from those found for both forms of imagery. Again, while this is not proof that response-bias is not the only source of the recalibration effect shown for imagery in these experiments, it does present strong evidence against such a claim.

## Conclusions

The experiments reported here extend the findings of Scott et al. (2013) by showing that not only can speech imagery influence concurrent speech perception but that speech imagery can have a lingering effect on perception. After repeated exposure to an ambiguous sound that has been shifted by speech imagery, perceivers' phoneme boundaries are recalibrated, at least temporarily.

Such recalibration has already been shown for illusions such as the Ganong effect and visual influences on speech perception, so it may not seem surprising that inner speech, which can cause an illusion similar to these effects, can also induce recalibration. However, when vision influences speech perception, the audio and video information are plausibly coming from a common source, whereas in the speech-imagery illusion the external sound and the imagined sound are not plausibly from a common source and so may not interact in the same was as is seen in other influences on speech perception. This possibility is raised by the finding that recalibration is sensitive to source reliability (Kraljic et al., 2008a). Despite this issue of conflicting sources, recalibration from speech imagery was successfully demonstrated.

In conclusion, the experiments reported here demonstrate that, in addition to written, lexical and visual information, speech imagery can induce recalibration. It is possible that this is due to a common underlying mechanism tied to forward models, however these experiments do not test that possibility directly and are intended as a proof of effect of this new recalibration inducer.

**Acknowledgments** I would like to thank my research assistant, Akeela Fatheen Abdul Gafoor, for her excellent work in running participants in this experiment.

## References

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification a mcgurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Borden, G.J., Harris, K.S., & Oliver, W. (1973). Oral feedback I: Variability of the effect of nerve-block anesthesia upon speech. *Journal of Phonetics*, *1*, 289–295.

Cooper, W.E., Billings, D., & Cole, R.A. (1976). Articulatory effects on speech perception: a second report. *Journal of Phonetics*, *4*, 219–232.

Cornsweet, T.N. (1962). The Staircase-Method in psychophysics. *The American Journal of Psychology*, *75*(3), 485–491.

Cutler, A., McQueen, J.M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In *Interspeech 2008* (pp. 2056-2056).

Eimas, P., & Corbit, J. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*, 99–109.

Eisner, F., & McQueen, J.M. (2005). The specificity of perceptual learning in speech processing. *Perception & psychophysics*, *67*(2), 224–238.

Eisner, F., & McQueen, J.M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953.

Fox, R.A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(4), 526–540.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

Heavey, C.L., & Hurlburt, R.T. (2008). The phenomena of inner experience. *Consciousness and cognition*, *17*(3), 798–810.

Kawahara, H., Masuda-katsuse, I., & de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*, 187–207.

Keetels, M., Schakel, L., Bonte, M., & Vroomen, J. (2016). Phonetic recalibration of speech by text. Attention, Perception, & Psychophysics, In Press.

Kraljic, T., Brennan, S.E., & Samuel, A.G. (2008a). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81.

Kraljic, T., & Samuel, A.G. (2005). Perceptual learning for speech: is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Kraljic, T., Samuel, A.G., & Brennan, S.E. (2008b). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*(4), 332–338.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Cambridge: Blackwell.

Lawrence, M.A. (2013). ez: Easy analysis and visualization of factorial experiments. R package version 4.2–2.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Massaro, D.W., Cohen, M.M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: an examination across languages. *Journal of Phonetics*, *21*, 445–478.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Merckelbach, H., & Ven, V.V.D. (2001). Another White Christmas: fantasy proneness and reports of 'hallucinatory experiences' in undergraduate students. *Journal of Behavior Therapy and Experimental Psychiatry*, *32*, 137–144.

Miall, R.C., & Wolpert, D.M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*(8), 1265–1279.

Norris, D., McQueen, J.M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Oppenheim, G.M., & Dell, G.S. (2010). The flexible abstractness of inner speech. *Cognition*, 1–53.

Peirce, J.W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10.

Pickering, M.J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–10.

Pickering, M.J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2014). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1–117.

Pitt, M.A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 1037.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reinisch, E., Wozny, D.R., Mitterer, H., & Holt, L.L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics*, *45*, 91–105.

Rosenblum, L.D., & Saldaña, H.M. (1992). Discrimination tests of visually influenced syllables. *Perception & psychophysics*, *52*(4), 461–473.

Saldaña, H.M., & Rosenblum, L.D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America*, *95*(6), 3658–3661.

Sams, M., Möttönen, R., & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research*, *23*(2-3), 429–435.

Samuel, A.G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218.

Scott, M. (2013a). Corollary discharge provides the sensory content of inner speech. *Psychological Science*, *24*(9), 1824–1830.

Scott, M., Yeung, H.H., Gick, B., & Werker, J.F. (2013). Inner speech captures the perception of external speech. *Journal of the Acoustical Society of America Express Letters*, *133*(4), 286–293.

Scott, M.A. (2013b). Selective adaptation and corollary discharge in mouthed speech. *Journal of the Phonetic Society of Japan*, *17*(3), 1–9.

Sjerps, M.J., & McQueen, J.M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 195–211.

Skipper, J.I., Nusbaum, H.C., & Small, S.L. (2006). Lending a helping hand to hearing: another motor theory of speech perception. In M.A. Arbib (Ed.) *Action to Language via the Mirror Neuron System, pp. 250–285*. Cambridge: Cambridge University Press.

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., & Small, S.L. (2007). Hearing Lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral cortex (New York, N.Y.: 1991)*, *17*(10), 2387–99.

Stevens, K.N. (2000). *Acoustic phonetics*. Cambride: MIT press.

Summerfield, Q., Bailey, P.J., & Erickson, D. (1980). A note on perceptuo-motor adaptation of speech. *Journal of Phonetics*, *8*, 491–499.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, *1*, 1–23.

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, *6*(314), 1–11.

van Linden, S., & Vroomen, J. (2008). Audiovisual speech recalibration in children. *Journal of child language*, *35*(4), 809–22.

Vroomen, J., & Baart, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition*, *110*(2), 254–259.

Vroomen, J., & Baart, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech*, *52*(2/3), 341–350.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*(1-4), 55–61.

Wilton, R.N. (1985). The recency effect in the perception of ambiguous figures. *Perception*, *14*(1), 53–61.