The neural control of volitional vocal production – from speech to identity, from social meaning to song.

Sophie K Scott

Abstract

The networks of cortical and sub-cortical fields that contribute to speech production have benefitted from many years of detailed study, and have been used as a framework for human volitional vocal production more generally. In this article I will argue that we need to consider speech production as an expression of the human voice in a more general sense. I will also argue that the neural control of the voice can and should be considered to be a flexible system, into which more right hemispheric networks are differentially recruited, based on the factors that are modulating vocal production. I will explore how this flexible network is recruited to express aspects of non-verbal information in the voice, such as identity and social traits. Finally, I will argue that we need to widen out the kinds of vocal behaviours that we explore, if we want to understand the neural underpinnings of the true range of sound making capabilities of the human voice.

Keywords

Speech, voice, vocal modulations, voluntary and involuntary vocalizations.

Introduction

The human voice is one of the most complex sounds in nature. At the heart of our acquisition of language, our voices are also musical instruments of extraordinary power and variety, and we are also excellent vocal mimics. Our voices are social acts, expressing both language and indexical personal information when we speak (Belin et al, 2004) – and we most commonly use our voices in social settings (Scott, Mcgettigan and Eisner, 2009). In terms of the neural basis of the production of human voices, we have two broadly distinct networks that underlie vocal behaviour – one largely associated with reactive, involuntary vocalizations (e.g. screams, swearing) (Jurgens, 2009) (Figure 1A) and one associated with complex, voluntary control of respiration, the larynx and the articulators (e.g. Jurgens, 2002;

Pisanski et al, 2016) (Figure 1B). The former is associated with a midline brainstem network, linking into the anterior cingulate, and is highly conserved across mammals (Jurgens, 2009). The latter, the volitional vocalization network (VVN), is a larger system including bilateral sensori-motor cortex, supplementary motor cortex, cerebellar fields, auditory cortex, sub cortical nuclei and left lateralized prefrontal and insular cortex (e.g. Blank et al, 2002).

In humans, the VVN has been most extensively explored in neuroimaging and clinical studies of speech production. This emphasis on speech and language is easily explicable, not least because damage to this system can lead to severe and pervasive problems with speech production. Indeed, early studies of speech production were clinical studies (Broca, 1861), and early functional imaging studies were often specifically guided by clinical findings (e.g. Wise et al, 1991). This has had the slightly biasing effect of the speech production network being treated as synonymous with the volitional vocalization network. In this paper I will argue that volitional vocal acts are much more varied than just speech, and speech expresses more than simply language.

The speech production network

The classic work of the 19th century neurologists established a critical role for the posterior third of the left inferior frontal gyrus in the production of speech. The most famous case was that of Paul Broca's patient Tan, who could only produce the word Tan (and some 'uncouth' swearing). A post-mortem analysis of his brain revealed that Tan had a tumour in the left posterior third of the left inferior frontal gyrus (Broca, 1861, Mohr, 1976, Dronkers et al, 2007). This apparent direct link of speech production to a specific brain region was the first cognitive faculty to be associated with some form of cortical specialization. However, further post-mortem work in the 1970's showed that cortical lesions which solely encompass the territory of 'Brocas area' are associated with transient mutism: to have a persistent problem with the production of speech the damage needs to extend into the surrounding white matter tracts (Mohr et al, 1978, Alexander et al, 1990). Indeed, recent research suggests that Broca's original patient did indeed have more extensive damage (Dronkers et al, 2007). Speech production issues associated with damage to Broca's area

could also be due to damaged connections to a wider cortical and subcortical network, and subsequent studies have revealed a wider neural system associated with speech production, while reaffirming an important role for the left posterior third of the inferior frontal gyrus (left pIFG) (e.g. Blank et al, 2002).

Functional imaging studies of overt speech production have shown that in addition to the left posterior third of the inferior frontal gyrus, speech production recruits a highly reliable network including bilateral primary motor cortex, bilateral primary sensory cortex, bilateral superior temporal gyri, supplementary motor cortex, bilateral paravermal cerebellum, left premotor cortex, left anterior insula and the, left posterior part of the supra-temporal plane, as well as the dorsal brainstem, pallidum and putamen (Wise et al, 1999, 2001; Blank et al 2002) (figure 1B, figure 2). Some of the functional roles of these regions in speech production may be hypothesized by their anatomy and physiology – for example, primary motor cortex contains a somatotopic map of the articulators (Penfield and Rasmussen, 1950; Bouchard et al, 2013), and neurons from here project directly to brainstem motoneurons, enabling fine control of the motor acts of speech. The supplementary motor cortex (SMA) is formed of two distinct fields – the more posterior SMA proper is densely connected to both motor cortex and to brainstem motor neurons, unlike the more anterior pre-SMA (Tanji, 1994). Functionally, SMA proper is strongly linked to movement generation and control, while pre-SMA is associated with action preparation and sequence planning (reviewed in Lima, Krishnan and Scott, 2016). Somatosensory and auditory fields are critical for the online sensory guidance of actions (e.g. Lametti et al, 2012), and both project into primary motor cortex. Premotor cortex may be important for the co-ordination of sequences of actions, while studies have implicated the anterior insula in the planning of articulations (Dronkers 1996; Wise et al, 1999). The left pIFG – the area Broca first identified– seems to play less of a role in articulation and more in a transformative role between perceptual and production networks (Flinker et al, 2015). The cerebellar and basal ganglia fields may be part of the motor loops between the premotor and motor cortex (Wise et al, 1999). Damage to many of these fields can result in speech production problems (such as aphasia, dysarthria, apraxia). The left posterior part of the temporal plane has been reported in many functional imaging studies speech production and silent

articulation, and has been identified as an important sensory-motor link (Hickok et al, 2000; Wise et al, 2001).

Do we only express speech when we speak?

However, a critical question is whether the speech production system is only responsible for speech? For example, when we produce volitional vocalizations, we are arguably expressing our identities as well as our words. Lavan and collegaues (2018) explored this by comparing people's ability to discriminate between the voices of unfamiliar talkers. These talkers were recorded producing volitional laughter and involuntary laughter. The logic was that if identity in the voice is solely a result of some inherent property of people's anatomy (height, vocal tract length etc), then discrimination of unfamiliar speakers would be the same across all vocalization classes, as they are all produced by the same vocal tract. If, however, the *volitional* nature of vocalizations being produced affect the unfamiliar vocal identity discrimination task, it would suggest that there is a role for the construction and controlled expression of identity in the volitional control of the voice. And this was indeed the case: listeners were most accurate at discriminating unfamiliar talkers when the unfamiliar voices were laughing in a controlled, volitional way, but not when the unfamiliar voices were laughing spontaneously. In other words, volitional laughter contained sufficient information to distinguish between unfamiliar talkers. In contrast, 'information about identity is less successfully encoded in spontaneous laughter vocalizations' (Lavan et al, 2018, pp144). This suggests that when we produce non-speech sounds in a volitional manner, that part of what we are expressing is contributing to our vocal identity. Is the same found for speech? Is more than linguistic information being expressed in these voluntary vocal acts ? What are the implications of this for the speech production network – is it controlling how we sound as well as what we say?

It's certainly the case that patients who have suffered a stroke and who have an expressive aphasia frequently complain that they no longer 'sound the same': even if they have made a good recovery in terms of the fluency of speech production, they do not sound the way they used to. This loss of vocal identity along with speech production has even led to aphasia

being described as 'identity theft' (Shadden, 2010). Perhaps the most striking effect of this is foreign accent syndrome (Blumstein et al, 1987), where people speak with what sounds like a markedly different accent following a head injury. It has been argued to represent a form of speech apraxia (Varley and Whiteside, 2001). In my anecdotal experience of working in foreign accent syndrome (Scott et al, 2006), patients often care a lot less that their speech is agrammatic and inaccurate (though these frequently form the target of speech and language therapy) than that their voice itself is so altered that their fellow countrymen think that they are not native speakers. Of course, there is no new 'accent' in foreign accent syndrome, it is in the ears of the listeners, who are trying to classify and understand the meaning of this different voice, but a critical point to note is that the language disorder is unavoidably affecting the spoken voice of the patients, in a way that affects their sense of vocal identity. One patient describes her altered voice:

'… I went through this door; when I came through it the next morning it was not me'. 'Where did I go to?' 'It's not me, you know, it's somebody else'. 'People ask me where I come from. I say S. They say, I never heard anyone in S talk like that before. I think: that's right. I come from here, but I don't come from here any more. Where do I come from, where did I go to?' (Miller et al, 2006)

The experience of people with foreign accent syndrome and expressive aphasia are a potent demonstration that personally important elements of vocal identity are often damaged alongside speech production skills.

Variation in the voice.

The reliability of the voluntary vocal control network, which can be easily seen in individual participants using overt speech tasks in fMRI (figure 2), masks a greater complexity. The network is reliable, but speaking voices are highly
flexible (Lavan et al, 2018) – how can this neural system account for this plasticity in sound production, and the factors that lead to this?

Figure 3 shows short sections of speech from the same female talker, in two different speaking conditions. The upper panel shows her producing speech for use in testing – the words are clear and slow, and the pitch is low. The lower panel shows the same talker, in a session where laughter was being recorded, asking for help in the anechoic chamber from a colleague. The pitch is much higher (possibly because she is audibly smiling (Juslin and Laukka, 2003)), the speech is much faster, and many of the speech sounds are reduced (e.g.

'*going to'* becomes '*gonna'*). This is just one glimpse of the highly variable nature of human voices. Indeed, it can be argued that exposure to this variability is, under some circumstances, useful for us to learn about new vocal identities (Lavan et al, 2019). There are many reasons for this variability. As voices are motor acts, they are open to modulations, from the physiological, involuntary and voluntary mechanisms that can impact upon the ways that these motor acts are performed, especially in social contexts, such as conversations (e.g. Pardo et al, 2017).

Involuntary influences - Physiology/Health. As different physical states lead to different kinds of physiological affects on the body, these have differential effects on the voice. For example – the physiology of fear responses involves a release of adrenaline, which can affect the vocal folds, to the extent that professional singers can find that their voices are noticeably altered by performance anxiety (Giddens et al, 2013). Illnesses that affect respiration or cause inflammation of the vocal folds will have a noticeable effect on the voice, and if lung cancers affect the laryngeal nerve, vocal hoarseness can be a presenting symptom (Feierabend and Shahram, 2009).

Involuntary influences – involuntary vocal network. In addition to the network for voluntary vocalization (VVM) described in the introduction, humans also retain the midline vocalization network that is associated with involuntary, reactive vocalizations in humans and other mammals (Jurgens, 2009) (Figure 1A). This is comprised of sensory and motor nuclei in the lower brainstem, which are connected and co-ordinated by the reticular formation. The reticular formation also contains a vocal pattern generator, and this interacts with and can be inhibited by the periaqueductal grey and the anterior cingulate cortex (Jurgens, 2009). In humans, this involuntary network is associated with the production of non-verbal emotional vocalizations, and highly automatized examples of speech, such as swear words. Patients with an expressive aphasia – like Broca's patient Tan – often have these kinds of vocalizations preserved (Van Lancker and Cummings, 1999). The reactive, automatic nature of these involuntary vocalizations means that voluntary vocal acts speech can be strongly affected by emotions – both by the physiological/emotional changes mentioned above, and by the emergence of involuntary non-verbal emotional vocalizations that can directly interact with voluntary control of the articulators, making

speech and song difficult or impossible (Mcgettigan and Scott, 2014). The sound of someone trying to talk while being overcome by persistent emotional states such as laughter or weeping is the sound of these two networks in direct competition, a competition that the involuntary vocalizations frequently win. Figure 4 shows the spectrogram of a news presenter on the BBC, apologizing for laughing whilst also laughing – the interjections of laughter are involuntary, and will be viewed dimly by the BBC.

Developmentally, humans progress from an ability to produce some reactive involuntary vocalizations (e.g. crying) soon after birth; others (such as laughter) start to appear at around 3 months of age, and shows some rapid progressions in complexity over the first year of life (Sroufe and Wunsch, 1972). As the voluntary vocalization network matures, children start to produce progressively more complex speech and vocal acts are produced (Simonyan et al, 2016). As in babies, non-human primates and apes show some ability to modulate involuntary nonverbal vocalizations during across their lifespan (Takahashi et al, 2015; Lameira et al, 2016). However their lack of the ability to learn to use more complex vocalizations may relate in part to a neural limit on the possibilities of their ability to engage more voluntary networks to control their respiration and articulation.

Social and environmental influences on speech

Speech itself is highly variable, depending on the social context, the nature of the relationships and affiliations of the people interacting, and what emotional register the interaction is taking place in (e.g. Lavan et al, 2018). Voices are primarily used in conversational speech, where people align their voices and their breathing to facilitate the conversational interaction (Garrod and Pickering, 2004, Mcfarland 2001), People will change their voice depending on who they are talking to (Pardo and Remez, 2006), and the apparent communicative needs that are present: people talk differently to babies and children than to other adults (e.g. Lavan et al, 2018). People will often change how they speak depending on how they feel about the person they are talking to – the more they like the person they are talking to, the more they will be likely to align their voice with theirs, in terms of pitch, rhythm and accent, word use and syntactic structures (Chartrand and Bargh, 1999). And although we typically study speech and voice in the lab using highly artificial stimuli, natural speech does contain errors, omissions and dysfluencies, though overt errors

are rare - running at an average of 1 speech error for every 900 words (Garnham et al, 1982). In spontaneous speech, only around 60% of errors are corrected (Nooteboom 1980). Speech will also be produced differently depending on the auditory environment in which it is produced – in the Lombard effect, people will (partly unconsciously) raise the level of their voice to compete with environmental noise (Lombard, 1911). Talkers have even been shown to modulate their speaking voices to exploit the acoustic characteristics of masking sounds (Lu and Cooke, 2008, Cooke and Lu, 2010). The next sections consider the evidence for neural systems underlying such variation.

Neural systems underling speech variation

Blank and colleagues (2002) contrasted free propositional speech (produced in response autobiographical questions) with counting aloud and reciting familiar nursery rhymes (Blank et al, 2002). This showed the expected motor networks associated with speech (common to all speaking conditions), and a wider semantic network associated with the contrast of the propositional speech over counting aloud and reciting nursery rhymes. An unexpected incidental finding was that the left medial supratemporal plane response, which was present (as expected) for all three speech production conditions, was notably greater for the highly overfamiliar and highly rhythmic nursery rhyme conditions. Other than the finding that the total amount of neural activity seen correlated positively with the total amount of speech produced (Wise et al, 1999), this was an early indication that the *kind* of speech act and the *manner* in it was produced could affect the ways that elements of the motor speech production network were recruited. More detailed studies, expressly addressing different kinds of vocal modulations and their neural bases, are examined here.

Vocal Impressions Direct and deliberate alterations of vocal speaking style were explored in a study that was designed with a professional vocal impressionist (Mcgettigan et al, 2013). In this study, participants (who were not professional impressionists) were asked to say a familiar phrase aloud when cued to speak (e.g. 'humpty dumpty sat on a wall, humpty dumpty had a great fall'). They were also told to speak (a) in their 'normal' voice, or (b) cued to attempt to produce the phrase with a specific regional accent of English (e.g. Welsh, or Liverpool), or (c) with a specific vocal identity (e.g. The Queen, Donald Duck) (participants prespecified the target accents/identities that they were happy to attempt). While all three

speaking conditions led to common areas of activation in the sensorimotor cortex, left premotor cortex, and bilateral auditory fields, there was significantly greater activity for the two vocal change conditions over speaking in 'normal' voice in the left pIFG and anterior insula. The contrast of specific impressions over regional accents revealed activation in bilateral posterior superior temporal gyri, extending up into the bilateral inferior parietal cortex, and running along the right mid-anterior STS. The involvement of the left pIFG and anterior insula suggests that these canonical regions associated with speech production are also critically important in the controlled production of indexical information in the voice, such as identity and regional accent, as well as the linguistic information. When specific vocal identities are being attempted, the recruitment of right STS fields may speak to the use of talker information, which has been linked to the right temporal lobe (Belin et al, 2004, Roswandowitz et al, 2018). Overall, this study suggests that the volitional motor system underlying the production of speech may also be important in the expression of indexical information, such as accent and identity.

Altered perceptual information during speech production Sensory information is critical to the production of intentional movements – many manual tasks are associated with visual guidance of action, and speech production depends heavily on auditory and somatosensory guidance of action (e.g. Lametti et al, 2012). This can be used to investigate the neural basis of vocal modulations, by altering the perceptual consequences of speaking aloud. This can include asking people to speak aloud while they hear their voice over headphones, shifted up or down in pitch, or delayed in time to different degrees, with an altered spectral profile, or masked by noise (as in the original Lombard effect) (e.g. Howell, 2004). This kind of altered auditory information is commonly referred to as feedback, though feedback is a somewhat complex term: it is sometimes used to refer to perceptual processing of the sensory information associated with the guidance of speaking, and sometimes to refer to explicit feedback of sensory information, which is used to detect and correct errors in speech (Howell, 2004). Nonetheless, when we speak and the sounds we make are at odds with our expectations, we will often change our voice in response to this change. What neural systems are associated with this variation? A recent meta-analysis, (Meekings and Scott, 2021) drew together papers which had used altered pitch, timing and noise masking in a variety of speech production tasks. The analysis revealed significant overlap between

foci in bilateral superior temporal gyri (STG), transverse temporal gyrus, and right precentral gyrus. In the STG, overlap was concentrated in lateral auditory cortex and was more widespread in the right than the left hemisphere. Across these altered speech production tasks, there is common bilateral auditory cortex activation, more extensive on the right than the left.

We do not only use sound to help us guide articulations. Anyone who has tried to speak after dental anesthesia will have experienced the difficulty associated with altered somatosensory information during speech production – speech is critically dependent on somatosensory information about the articulators, which is normally highly correlated with the auditory information. The neural basis of adaptation to altered somatosensory information has been studied by Ostry and colleagues (Darainy et al, 2019). They used robotic devices to alter the dynamic movements of the articulators during speech production, and shown both patterns of adaptation and of altered perceptual processing of speech sounds following such interventions. The first part of the study used fMRI of normal speech production and perception to establish a 'listening and repeating' network. This was used to generate seed voxels for a connectivity analysis of resting state fMRI data. These resting state data were collected before and after a sensori-motor adaptation speech production task using the robotic arm. The robotic arm both collected data about the jaw movements and applied forces to the lower jaw, while they were prompted to produce the words *head, said, ted, bed* – these were chosen as jaw movements could alter the vowel's formants – for example, *head* could be produced or heard as *had*. The design enabled them to determine changes in resting state connectivity that were associated with these adaptation processes.

Their analysis of the changes in resting state connectivity for seed voxels chosen from the repetition task showed connectivity changes associated with motor changes in the sensori-motor adaptation task, which were dissociable from connectivity changes associated with the perceptual adaptations. Changes to articulation associated with the adaptation task were associated with connectivity increases between STG and somatosensory cortex, and between pre-SMA and right inferior parietal cortex.

The social voice The production of speech is typically considered to be a form of communication but inherent in this is the idea that we are communicating *with someone* – that is, the act of speaking is a social act. As noted earlier, the social context that we are in will greatly affect how our voice sounds: we have previously argued that speech production really should be considered to be a social behaviour (Scott, Mcgettigan and Eisner, 2009). This has been explicitly addressed in a couple of studies.

Mcgettigan and colleagues (Guldner et al, 2020) asked participants to overtly express two different dimensions of social information – participants were asked to read non-words aloud and to express social traits of competence (e.g. intelligent) and affiliation (e.g. hostile, likeable). The baseline condition was to use their normal voice, and the control for non-social vocal change was to sound larger. Any vocal changes (social traits and size over the baseline 'normal' voice) were associated with increased activation in bilateral anterior insulae, right STG, left IFG, the SMA extending into the anterior cingulate and the left supramarginal gyrus (SMG).

In contrast the expression of social traits (over the non-social trait of size) resulted in activations in memory related circuits such as the left hippocampus and bilateral retrosplenial, visual imagery related fields such as the left cuneus and precuneus and the bilateral lingual gyri, and semantic fields such as the medial prefrontal cortex and bilateral anterior STS. These regions have all be argued to fall within the social brain network (Guldner et al, 2020).

This is an important study, showing how elements of the social brain network are interacting with voluntary voice modulation networks to effect vocal change. But what would happen if we did not require people to deliberately change their voices, but gave them a social task where they would have to change their voice to align it with that of another person? One such task is joint speech, where two people reading from the same text (or reciting the same familiar text). Common in all human communities, from prayers and worship, to allegiance and pledges, joint speech is an interesting example of the extent to which people can accurately align their voices – people are capable of very tight temporal alignments, partly due to their converging on highly stereotypical patterns of vocal melody

and rhythm, and partly due to their aligning their breathing, and partly due to their each paying attention to the other. As many of these features are seen during conversational speech (Garrod and Pickering, 2004), joint speech is both an interesting paradigm for exploring vocal alignment, and joint action in general, and also for exploring some of the same alignments that occur in conversational speech. In an fMRI study (Jasmin et al, 2016), we compared people speaking aloud, listening to speech, performing joint speech with an experimenter in the control room, and performing joint speech with a recording of the experimenter (the participants did not know that this was two different conditions). The baseline was rest.

The two joint speech conditions, over listening/speaking alone, led to increased activity in the supratemporal plane and the superior temporal gyrus (STG), extending into the right inferior parietal cortex. When the two joint speech conditions were individually compared to listening and speaking alone, the live joint speech condition led to significant activity in the right inferior frontal gyrus, which was not present for the recorded joint speech condition. When the two joint speech conditions were directly compared with each other, the right IFG, right supramarginal gyrus and angular gyrus, right temporal pole, and bilateral parahippocampal gyri were all more activated by joint speech with a live speaker. As speakers were not aware that there two joint speech conditions were different, these differences are assumed to be outwith awareness, and may reflect processes associated with the coherence of the two voices and consequent greater accuracy (not possible when one of the voices is a recording). A PPI analysis of the right IFG and the right STG peak responses converged in common activation in the right somatosensory-motor cortex.

Modulations of the voluntary vocalization network and vocal change

These studies all show both the core voluntary vocal control network and regions beyond this are recruited to support vocal change, with an intriguing emphasis on right hemispheric involvement across many of the studies. The voluntary vocal control network has many bilateral elements (sensorimotor cortex, cerebellum, superior temporal gyri, basal ganglia), and where there are lateralized elements, these are in the left hemisphere. Is the recruitment of right hemispheric mechanisms critical for supporting variation in vocal production? There is also an interesting task variation – some of these tasks require people

to explicitly vocalize differently, while others give a task (e.g. altered perceptual consequences of speaking, joint speech) that will cause people to adjust their voice without explicit instruction.

The next section will address variation in the voluntary control of the voice by exploring the neural basis of wider kinds of vocal performance – instead of variations in the speaking voice, I will consider a wider range of vocal performances. Can we see systematic variation in the voluntary vocalization network when we move into song, rap and beat boxing?

Beyond speech: singing, vocal performance, and expertise

The vast majority of papers on the voluntary vocalization network are studying speech: this is of course a completely understandable bias, given the importance of speech for social communication, and the limitations that a problem with speech production can place on someone's life. However, this does lead us to forget the wealth of other vocal skills that humans can perform with their voices. What is the neural basis for these different kinds of vocal production?

Singing

Jeffries et al (2003) showed that singing the lyrics of songs, relative to speaking them, lead to a significantly greater pattern of activation in a right lateralized network, including the right dorsolateral prefrontal cortex, right secondary somatosensory cortex, right temporal lobe fields and the nucleus accumbens (Jeffires et al, 2003). This finding has been widely replicated in studies of speech and singing – while speech and song share many aspects of the wider voluntary vocalization network (Zarate, 2013), singing frequently recruits more right hemisphere regions than speech, when speech and melody generation networks are compared to rest (Brown et al, 2006), and when singing is directly contrasted with speaking (Özdemir et al, 2006) (reviewed in Mavridis and Pyrgelis 2016). Why is this? The right hemisphere shows a clear specialization for the perception of pitch changes and melody (reviewed in McGettigan and Scott, 2012, Scott and McGettigan, 2013) – are these mechanisms necessary to guide production where pitch change and melody are critical? This cannot be the whole story, however. Singing has been frequently shown to recruit right frontal fields, including the right homologue of Broca's area (Brown et al, 2006; Özdemir et al, 2006), which speaks to more than a pure role for perceptual guidance for the right

hemisphere. Indeed, at least one study has found that during 'normal' speech production, right IFG is actively suppressed (Blank et al, 2003), which implies a complex dynamic between the left and right IFG during speech production. What might this mean for vocal modulations?

Damage to the left IFG typically leads to considerable speech production issues, however some recover of speech production is common. Blank and colleagues (2003) demonstrated that right pIFG is recruited to support speech production in these cases – that is, plasticity in seems to be driven by changes in the right homologue of Broca's area. Notably, prior to any marked recovery of speech, people who are experiencing an ongoing period of expressive aphasia following left pIFG damage often can still sing – for example, they can sing sentences that they might struggle to speak. This retention of song in the context of profound problems with the production of speech forms the basis for melodic intonation therapy, which explicitly uses the melody and rhythm of music to rehabilitate speech production skills (e.g. Wilson et al, 2006). Patients expressive aphasia following damage to left pIFG can also typically perform joint speech tasks (Fridriksson et al, 2012), which Jasmin et al (2016) showed was associated with right pIFG activation. Perhaps right pIFG is engaged, by a process of alignment or entrainment to a melody or another speaker, to a degree that is not encountered in solo speech. Or perhaps right pIFG is recruited when significant deviations from a normal or solo speech production mode are required. One intriguing study of intraoperative electrostimulation of an opera singer during awake surgery for a right fronto-temporal glioma (Herbert et al, 2015), showed that stimulation of the right ventral premotor cortex led to severe dysarthria or speech arrest, but stimulating the right pars opercularis led to a switch from speech to song during naming, mentalising and spontaneous speech. The authors linked this to the many years of training that the singer had undergone, and this may well be highly relevant, but it also suggests that right pIFG is important in the controlled production of song, and that this may be why it is suppressed during 'normal' speech.

Different types of singing

I have contrasted speech and song fairly simplistically so far. However, while singing is a human universal behaviour, there are very many different styles of song, some of which are

associated with extensive training, and some genuinely extraordinary vocal feats. Many singers develop through training a 'singer's formant', which enables their voice to stand out from the ongoing acoustic context (e.g. an orchestra) (Sundberg, 2001). This training can also include learning to use the singing voice with a great deal of power, to extend their vocal range, and other techniques such as vibrato.

A study (Kleber et al, 2009) compared the neural activation during singing (contrasted with silent breathing) in trained opera singers, conservatory level singing students, and people with no singing training. They reported common activations during song in a network including the bilateral STG, bilateral sensorimotor cortex, SMA, and right inferior premotor cortex. The professional opera singers showed greater activation than both groups of non-opera singers in right primary sensory and motor cortex, the precuneus, the putamen and cerebellum. Kleber and colleagues (2016) next used voxel based morphometry to identify experience dependent effects of opera singing training on brain structure, and found that a subset of these functional brain differences were also showing increased volumes compared to controls. Opera singers showed right hemisphere cortical volume increases in primary somatosensory fields, associated with articulatory and laryngeal representations, which extended into the supramarginal gyrus, primary auditory cortex and ventral secondary somatosensory cortex. These are a powerful demonstrations of the ways that extensive training can shape cortical processing, and again it is striking that the cortical effects are right lateralized. It is also notable that pitch imitation in humans is strongly linked to the putamen, orofacial sensorimotor cortex and the SMA, so elements of this skill may be associated with the pattern seen in opera singers (Belyk et al, 2016).

Rapping

Rapping is a musical vocal style that emphasizes rhyme, rhythmic speech, and street vernacular. Often (though not exclusively) associated with improvisation, rapping is also an extremely interesting instrumental style, with very complex rhythmic patterning. Rapping can also be extremely fast – the world record is held by Twista, who achieved a rate of 11.2 syllables per second (that is, an average syllable duration of 89.3ms). There is also some evidence that the singer's formant (Sundberg, 2001) is also exploited by rappers, who employ this technique when producing syllables on the beat, but not off the beat (when it is less likely to be masked by the drum beats) (Ammirante and Copelli, 2019). Another study

suggested that there are different vocal styles within the wider genre of rap some of which lean further towards song, others towards vernacular speech (Ohriner, 2019). The one fMRI study of the neural basis of rapping (Liu et al, 2012) emphasized the improvisation element, contrasting this with well rehearsed rapping sequences: this revealed greater activation for improvisation in medial prefrontal cortex and left dorsolateral prefrontal cortex, while the familiar, pre-rehearsed rap sequences led to greater activation in the right dorsolateral prefrontal cortex. I suspect that further investigation of this kind of vocal expertise will enable us to unpack the contributions of vocal control to this pattern of results.

Beat boxing

Beat boxing, which has likely been around for millennia, has its roots in the use of the human voice to produce percussive sequences. Beat boxing has become more popular in the last forty years due initially to its link to rap and hip-hop (where beat boxers would provide a beat for rappers by mimicking drum machines), and then as its own musical form. Beat boxing as it exists today rarely consists of a simple rhythm track, and beat boxers typically produce complex polyphonic musical sequences, by using a wide range of articulations at a very fast rate, and by exploiting parallel ways of making vocal sounds (e.g. vocal fold vibrations, nasal harmonics, bilabial vibrations). Like rapping, beat boxing is not a skill people generally learn in a formal musical training, and because of the complexity and flexibility of the human articulators and neural control of the articulators, respiration and larynx, different beatboxers can have extremely different techniques. Current functional imaging studies have emphasized the ways that this skill shapes the perception of musical sequences (Krishnan et al, 2017), and dynamic MRI of the vocal tract has been used to explore the techniques that beat boxers use to achieve this fast and complex vocal repertoire (Proctor et al, 2013). Further studies exploring the basis for the neural control of these sequences of sound will be able to determine the extent to which this can be attributed to the core volitional voice control system, and how this is supplemented by other neural systems.

Conclusions

If science is a map that we develop to explore our worlds, then our scientific map of the voice has a great deal of detail about some aspects of the neural basis of speaking aloud,

and some vast unmarked territories that are either unexplored, or not yet integrated into the map. There is much still that we do not know. There are no studies of many different kinds of speaking styles (e.g. rhetoric, poetry, comedy, acting, expert vocal impressions) all of which signal a different kind of engagement with the audience: nor are there any studies explicitly addressing human vocal mimicry – we know a great deal more about this skill in non-human animals than we do in humans (Fitch, 2000). The technical difficulties of studying vocal production in fMRI mean that progress is slow (compared to studies of perception), but techniques improve and we are already asking wider questions about the neural control of the voice than we were doing twenty years ago.

However, we can see that variability in the human voice can be associated with activity in the core volitional voice control network (summarized in Table 1). The left pIFG is recruited for speech production, but also for achieving different accent and identity targets, and different social and non-social vocal cues. Auditory fields are recruited to adapt to changes in the auditory outcomes of produced speech, and somatosensory cortex shows connecting changes associated with adaptations to altered articulator dynamics during speech production. This network is also complemented by right lateralized systems. Right pIFG is recruited during joint speech and song, right ventral motor cortex is implicated in adaptations to speaking under different auditory contexts, opera singers recruit right sensorimotor cortex more than other singers, and right STS fields are recruited when people attempt to sound like specific individuals. Another difference that may be important in these studies is that some instruct the participants to actively change their voices (e.g. the vocal changes used in Mcgettigan et al, 2013, Guldner et al, 2020), while others give the participants tasks that will implicitly lead to vocal change, even if the participants are not directly aware of this (e.g. the auditory manipulations of speech production discussed by Meekings and Scott, 2021, the joint speech task used by Jasmin et al (2016)). What are the implications of these different kinds of task requirements on the patterns of neural activations seen in the volitional vocalization network?

A better understanding of this wider voice network, its dynamics and its connectivity, will enable a more rigorous exploration of the computational properties of this flexible volitional vocal control system, and how it interacts with other neural systems. Part of this is going to

require us to ask more questions about the nature of these hemispheric differences, and how they contribute to vocal production, and part of this will require us to ask questions about the human voice in a way that lets us benefit from and explore its flexibility and range. Speech is a great model system for exploring these networks, but speech is highly flexible and variable, and still reflects only a subset of human vocal ability in its wider sense.

References

Alexander MP, Naeser MA, Palumbo C. (1990) Broca's area aphasias: aphasia after lesions including the frontal operculum. Neurology;40:353–362.

Ammirante, P; Copelli, F (2019) Vowel Formant Structure Predicts Metric Position in Hip-Hop Lyrics, Music Perception, Volume: 36 (5): 480-487

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. Trends in Cognitive Sciences, 8, 129–135.

Blank SC, Bird H, Turkheimer F, Wise, RJS (2003) Speech production after stroke: The role of the right pars opercularis. Annals Of Neurology Volume: 54 (3): 310-320

Blank SC, Scott SK, Murphy K, Warburton E, Wise RJ. (2002) Speech production: Wernicke, Broca and beyond. Brain; 125(8):1829-38.

Blumstein SE, Alexander MP, Ryalls JH, Katz W, Dworetzky B (1987) On the Nature of the Foreign Accent Syndrome - A Case-Study. Brain and Language, 31(2): 215-244

Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation, Nature 495:327–332.

Broca P (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). Bulletin de la Société Anatomique, 6, 330–357.

Brown S, Martinez MJ, Parsons LM (2006) Music and language side by side in the brain: a PET study of the generation of melodies and sentences. European Journal of Neuroscience, 23: 2791–2803.

Chartrand TL, Bargh JA (1999). The chameleon effect: The perception–behavior link and social interaction. Journal of Personality and Social Psychology, 76, 893–910.

Child Development, 43 (4): 1326-1344

Cooke M, Lu Y (2010). "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," J. Acoust. Soc. Am. 128(4), 2059–2069.

Darainy M, Vahdat S, Ostry DJ (2019) Neural Basis of Sensorimotor Plasticity in Speech Motor Adaptation. Cerebral Cortex, 29(7): 2876–2889.

Dronkers NF, Plaisant O, Iba-Zizen MT, Cabanis EA (2007). Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong, Brain, 130(5):1432–1441

Dronkers NF. A new brain region for coordinating speech articulation. Nature 1996;384:159–161.

Feierabend RH, Shahram MN (2009) Hoarseness in adults. American Family Physician, 80(4):363-70.

Fitch WT (2000) The evolution of speech: a comparative review. Trends in Cognitive Sciences 4(7): 258-267

Flinker A, Korzeniewska A, Shestyuk AY, Franaszczuk PJ, Dronkers NF, Knight RT, Crone NE (2015) Redefining the role of Broca's area in speech, PNAS 112 (9): 2871-2875

Fridriksson J, Hubbard HI, Hudspeth SG, Holland AL, Bonilha L, Fromm D, Rorden C (2012) Speech entrainment enables patients with Broca's aphasia to produce fluent speech. Brain. 135(12): 3815–3829.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitive Sciences, 8, 8–11.

Giddens CL, Barron KW, Byrd-Craven J, Clark KF, Winter AS (2013) Vocal Indices of Stress: A Review. Journal of Voice, 27(3): 390.e21-390.e29

Guldner S, Nees F, McGettigan C (2020) Vocomotor and Social Brain Networks Work Together to Express Social Traits in Voices. Cereb Cortex, 30(11):6004-6020.

Hickok G, Erhard P, Kassubek J, Helms-Tillery AK, Naeve-Velguth S, Strupp JP, Strick PL, Ugurbil K (2000) A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: implications for the explanation of conduction aphasia. Neurosci Lett. Jun 23;287(2):156-60.

Howell P (2004) Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering. Stammering Research, 1(1):1-31.

Jasmin KM, McGettigan C, Agnew ZK, Lavan N, Josephs O, Cummins F, Scott SK (2016) Cohesion and joint speech – right hemisphere contributions to synchronized vocal production. Journal of Neuroscience, 36(17):4669-80

Jeffries KJ, Braun AR, Fritz JB (2003). Words in melody: an H 2 15 O PET study of brain activation during singing and speaking. Neuroreport 14, 749–754.

Jürgens U (2002) Neural pathways underlying vocal control Neurosci. Biobehav. Rev., 26 (2002), pp. 235-258

Jürgens U (2009) The neural control of vocalization in mammals: a review J. Voice, 23 (2009), pp. 1-10

Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull., 129:770-814

Kleber B, Veit R, Birbaumer N, Gruzelier J, Lotze M (2010) The brain of opera singers: experience-dependent changes in functional activation Cereb Cortex, 20(5):1144-52

Kleber B, Veit R, Moll CV, Gaser C , Birbaumer N, Lotze M (2016) Voxel-based morphometry in opera singers: Increased gray-matter volume in right somatosensory and auditory cortices. Neuroimage, 133:477-483

Lameira AR, Hardus ME, Mielke A, Wich SA, Shumaker RW (2016) Vocal fold control beyond the species-specific repertoire in an orangutan. Sci Rep 6:30315

Lametti DR, Nasir SM, Ostry DJ (2012) Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback. Journal of Neuroscience 32 (27): 9351-9358.

Lavan, N Burton, AM Scott, SK McGettigan, C (2019) Flexible voices: Identity perception from variable vocal signals Psychonomic Bulletin & Review 26(1): 90-102

Lava, N, Short B, Wilding A, McGettigan, C. (2018). Impoverished encoding of speaker identity in spontaneous laughter. *Evolution and Human Behavior*, *39*(1), 139-145.

Lima C, Krishnan S, Scott SK (2016) Roles of Supplementary Motor Areas in Auditory Processing and Auditory Imagery. Trends in Neurosciences, 39(8):527-42.

Liu S , Chow HM, Xu Y, Erkkinen MG, Swett KE, Eagle MW, Rizik-Baer DA, Braun AR (2012) Neural correlates of lyrical improvisation: an FMRI study of freestyle rap. Sci Rep, 012;2:834.

Lombard, E. (1911). “Le signe de l’elevation de la voix” (“The sign of the elevation of the voice”), Annales Des Maladies de L’Oreille et Du Larynx 37, 101–119.

Lu Y, Cooke M (2008). “Speech production modifications produced by competing talkers, babble, and stationary noise,” J. Acoust. Soc. Am. 124(5), 3261–3275.

Mavridis IN, Pyrgelis ES (2016) Brain Activation During Singing: "Clef de Sol Activation" Is the "Concert" of the Human Brain. Med Probl Perform Art. 31(1):45-50.

McFarland DH (2001). Respiratory markers of conversational interaction. Journal of Speech, Language, and Hearing Research, 44, 128–143.

McGettigan C, Scott SK (2012) Cortical asymmetries in speech perception: what’s wrong, what’s right, and what’s left? Trends in Cognitive Sciences, 6(5):269-76

McGettigan C, Scott SK (2014) Voluntary and involuntary processes affect the production of verbal and nonverbal signals by the human voice. Behavioral and Brain Sciences, 37(6)

Miller N, Lowit A, O'Sullivan H (2006) What makes acquired foreign accent syndrome foreign? Journal Of Neurolinguistics, 19 (5):385-409.

Mohr JP, Pessin MS, Finkelstein S, Funkenstein HH, Duncan GW, Davis KR. (1978) Broca aphasia: pathologic and clinical. Neurology; 28: 311-24.

Mohr JP. (1976) Broca's area and Broca's aphasia. Vol 1. New York: Academic Press .

Nooteboom SG (1980) Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V.A. Fromkin (Ed.), Errors in linguistic performance: Slips of the tongue, ear, pen and hand, Academic Press, New York (1980), pp. :201–236.

Ohriner, M (2019) Analysing the pitch content of the rapping voice, Journal Of New Music Research,48(5): 413-433

Özdemir E, Norton A, Schlaug G (2006) Shared and distinct neural correlates of singing and speaking. NeuroImage 33:628–635.

Pardo JS, Urmanche A, Wilman S, Weiner J (2017) Phonetic convergence across multiple measures and model talkers. Atten Percept Psychophys 79, 637–659 (2017)

Pardo, J. S., & Remez, R. E. (2006). The perception of speech. In M. Traxler & M. A. Gernsbacher (Eds.), The handbook of psycholinguistics (2nd ed., pp. 201–248).

Pisanski K , Cartei V , McGettigan C , Raine J , Reby D (2016) Voice Modulation: A Window into the Origins of Human Vocal Control? Trends Cogn Sci . 20(4):304-318

Proctor M, Bresch E, Byrd D, Nayak K, Narayanan S. (2013) Paralinguistic mechanisms of production in human "beatboxing": a real-time magnetic resonance imaging study. J Acoust Soc Am. 133(2):1043-54.

Roswandowitz C, Kappes C, Obrig H, von Kriegstein K (2018) Obligatory and facultative brain regions for voice-identity recognition, *Brain*, 141(1):234–247

Scott SK, Clegg F, Rudge P, Burgess PW (2006) Foreign accent syndrome, speech rhythm and the functional neuronatomy of speech production. Journal of Neurolinguistics, 19(5): 370-384

Scott SK, McGettigan C and Eisner F (2009) A little more conversation, a little less action: candidate roles for motor cortex in speech perception. Nature Reviews Neuroscience. 10(4):295-302

Scott SK, McGettigan C. (2013) Do temporal processes underlie left hemisphere dominance in speech perception? Brain Lang. 2013 Oct;127(1):36-45.

Shadden B (2010) Aphasia as identity theft: Theory and practice Aphasiology, 19(3-5): 211-223

Simonyan K, Ackermann K, Chang EF, Greenlee JD (2016) New Developments in Understanding the Complexity of Human Speech Production. Journal of Neuroscience, 36(45): 11440-11448.

Sroufe LA, Wunsch JP (1972) The Development of Laughter in the First Year of Life. Child Development, 43 (4): 1326-1344

Sundberg J (2001). Level and center frequency of the singer's formant. Journal of Voice, 15, 176–186.

Takahashi DY, Fenley AR, Teramoto Y, Narayanan DZ, Borjon JI, Holmes P, Ghazanfar AA (2015) Language Development: the developmental dynamics of marmoset monkey vocal production. Science 349:734–738

Tanji J (1994) The supplementary motor area in the cerebral cortex, Neuroscience Research, 19(3):251-268.

Tremblay P, Deschamps I, Gracco VL (2016) Chapter 59 - Neurobiology of Speech Production: A Motor Control Perspective, Editor(s): Gregory Hickok, Steven L. Small, Neurobiology of Language, Academic Press, Pages 741-750.

Van Lancker D, Cummings JL (1999) Expletives: neurolinguistic and neurobehavioral perspectives on swearing. Brain Research Reviews, 31(1): 83-104

Varley, R Whiteside, SP (2001) What is the underlying impairment in acquired apraxia of speech? Aphasiology. 15(1): 39-49.

Penfield W, Rasmussen T (1950) The Cerebral Cortex of Man: A Clinical Study of Localization of Function, Macmillan.

Wilson SJ, Parsons K and Reutens DC (2006) Preserved Singing in Aphasia: A Case Study of the Efficacy of Melodic Intonation Therapy. Music Perception: An Interdisciplinary Journal, 24(1):23-36

Wise R, Chollet F, Hadar U, Friston K, Hoffner E, Frackowiak R (1991) Distribution of Cortical Neural Networks Involved in Word Comprehension and Word Retrieval. Brain, 114(4): 1803-1817

Wise RJ, Greene J, Buchel C, Scott SK. (1999) Brain regions involved in articulation. Lancet;353:1057–1061.

Wise, RJS, Scott, S. K., Blank, S. C, Mummery, CJ, Warburton, E (2001) Identifying separate neural sub-systems within 'Wernicke's area', Brain, 124, 83-95.
Zarate JM (2013) The neural control of singing. Front Hum Neurosci. 2013 Jun 3;7:237.

Figure 1 simplified diagrams of the involuntary vocalization network (A) and the speech production network (B) from Jurgens, 2002; 2009, Tremblay et al, 2016

Figure 2 example data from one participant speaking aloud in fMRI (data collected for a demonstration by Zarinah Agnew). The commonly noted cortical fields associated with speech production are 1. Left posterior inferior frontal gyrus, 2. left anterior insula, 3. left central premotor cortex, 4. Bilateral primary sensory-motor cortex (pre and post central gyri), 5. Bilateral superior temporal gyri, 6. Left posterior part of the temporal plane, 7. Supplementary motor area.

Figure 3 [A] spectrogram of a female adult talker, reading a sentence aloud for use in testing. Average duration of each syllable – 374 ms.

[B] spectrogram of the same female adult talker, talking spontaneously to colleagues during a recording session, where laughter stimuli were being recorded. Average duration of each syllable – 140ms.

Figure 4 spectrogram for adult male Scottish speaker, broadcasting live on BBC Radio 4. He is apologizing for laughing, and the laughs can be seen as abrupt increases in pitch.

Table 1 a summary of cortical fields in the right and left hemispheres, and the patterns of activations that they show, associated with different vocal tasks. Note that 'speech production' fields in the left hemisphere are associated with speech, and also with other kinds of vocal production changes (e.g. talker accent and identity). Note also that right hemisphere fields are differentially recruited by a variety of vocal tasks, from joint speech to song. LIFG Left inferior frontal gyrus, LAnt I left inferior frontal gyrus, LvPrem left ventral premotor cortex, LM1 left primary motor cortex, LS left somatosensory cortex, LAr left auditory cortex (rostral), LAc left auditory cortex (caudal), RIFG right inferior frontal gyrus, RAnt I right inferior frontal gyrus, RvPrem right ventral premotor cortex, RM1 right primary motor cortex, RS left somatosensory cortex, RAr right auditory cortex (rostral), RAc right auditory cortex (caudal).

Table 1

task	Speaking>rest	Vocal change Accent/identity	Change of identity	Altered auditory /somatosensory input	Joint speech	Song>speech	Opera
Cortical field							
LIFG	increase	increase					
LAnt I	increase	increase					
LvPrem	increase						
LM1	increase						
LS	increase						
LAr				increase	increase		
LAc	increase		increase	increase	increase		
RIFG	suppressed	increase			increase	increased	
RAnt I		increase				increased	
RvPrem						increased	
RM1	increase			increase			increase
RS	increase					increased	increase
RAr			increase	increase	increase	increased	
RAc			increase	increase	increase		

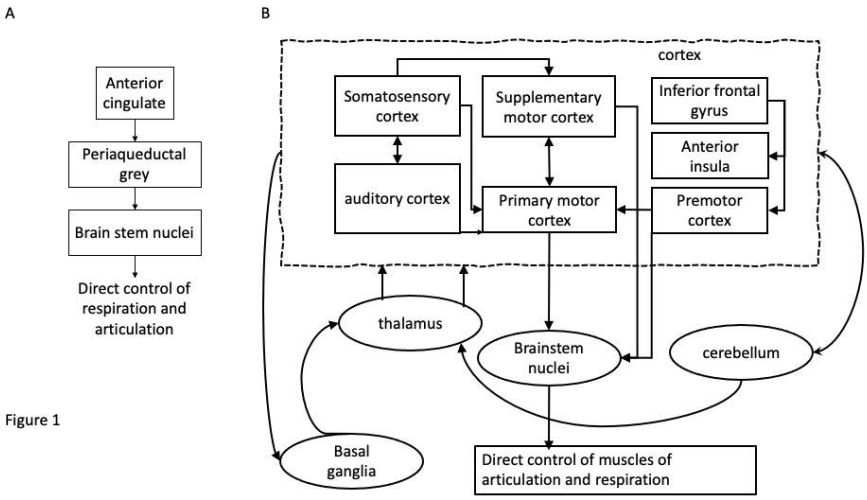

Figure 1

Figure 2

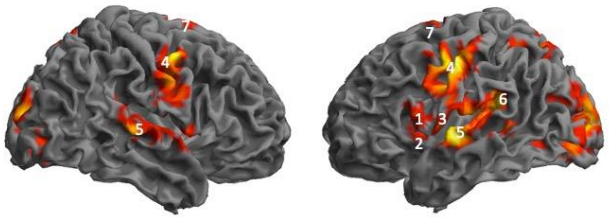

Figure 3

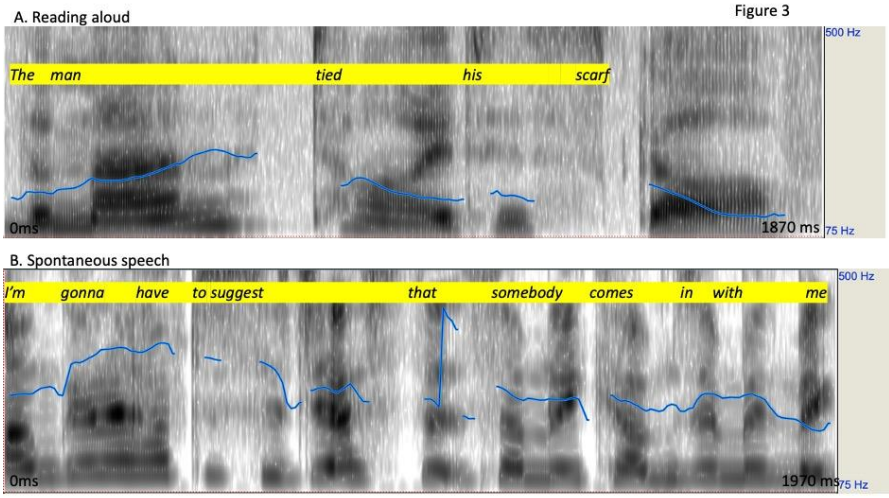

Figure 4