Speaker Normalization in Speech Perception

Keith Johnson
Matthias Sjerps

1 Introduction

Talkers differ from each other in a great many ways. Some of the difference is in the choice of linguistic variants for particular words, as immortalized in the song by George and Ira Gershwin “Let’s call the whole thing off”.

> You say either [iðɚ] and I say either [aɪðɚ],
> You say neither [niðɚ] and I say neither [naɪðɚ]
> Either[iðɚ], either [aɪðɚ] Neither [niðɚ], neither [naɪðɚ]
> Let's call the whole thing off.
>
> You like potato and I like potahto
> You like tomato and I like tomahto
> Potato, potahto, Tomato, tomahto.
> Let's call the whole thing off

Listeners have experienced different pronunciations of words, and many of the variants that we know are tinged with social or personal nuance. This “multiple-listing” notion, that listeners store more than one variant of each word in memory is the dominant hypothesis, among sociolinguists regarding the cognitive representation of social phonetic variation (Thomas, 2011), and has been proposed as a way to account for the listeners’s ability to ‘normalize’ for talker differences in speech perception (Johnson, 1997).

Beyond having experience and associations with particular variants of words, though, listeners are tolerant of unfamiliar variation. Un-experienced variants can nonetheless be recognized. It was common to experience this in the early days of text to speech synthesis when speech synthesizers could be counted on to pronounce some words in totally novel ways. For example, the hand-tuned orthography-to-pronunciation rules of a synthesizer would incorrectly pronounce *San Jose* as [sæn ʤoz], on analogy with *hose* (Liberman & Church, 1992). Listeners can also be exposed to previously unexperienced variations when listening to an unfamiliar dialect, and with a little exposure be able to cope with a new speech pattern. Interestingly, perceptual learning of such variation is pretty rapid (Greenspan et al., 1988). Listeners can use semantic context to guess the identity of a word even though its pronunciation is unfamiliar, and rapidly develop the ability to recognize new pronunciation variants.

Both of these processes -- multiple listing of variants, and top-down parsing of the speech stream -- can be seen as mechanisms to lend coherence to speech in the face of linguistic/phonetic variation. One main focus of this review is to consider whether there are aspects of auditory processing that help remove talker differences before the signal enters a multi-listing/top-down guided word recognition system.

Our conclusion, will be that the answer is ‘yes’ in two ways. First, auditory spectral analysis and encoding removes some talker differences. And second, contrast coding in an

auditory/phonetic frame of reference seems to apply before lexical processing begins. However, we will find that these mechanisms are partial, and that there is evidence both from behavioral studies and from neuro-imaging that indicates a role for expectation-guided coherence-lending mechanisms in speech perception.

2 Physiological/acoustic differences between talkers

Disregarding differences between talkers that may be due to different habits of articulation -- those talker differences that may be due to differences in dialect or social group, or even idiosyncratic habits (i.e. instances pronounced with the same linguistic variants), people have different acoustic voice 'signatures'. So much so, that voice can be used in biometric identification (Nott, 2018), and listeners can recognize familiar talkers (Hollien, 2001).

The largest acoustic difference between talkers is the difference between men and women and children. The physiological property that underlies this is the larger size, and lower location of the larynx in the neck in men (Fitch & Giedd, 1999). The larger size of the larynx is accompanied by longer and thicker vocal folds in men, and thus the mean rate of vibration of the vocal folds in voiced speech is lower in men than in women or children. The lower position of the larynx in the neck results in a longer vocal tract (and proportionally longer pharynx), and thus the resonant frequencies of the vocal tract (the vowel formant frequencies) are generally lower in men than in women. There is evidence (Dabbs & Mallinger, 1999; Zimman, 2017) of a tie between these voice features and testosterone, though this probably depends on the testosterone level (Glaser, et al., 2016).

Even within gender, vocal tract length is recoverable from the acoustic signal. Lammert & Narayanan (2015) found that a multiple regression formula combining the first four vowel formant frequencies can predict the measured length of the vocal tract to within about a centimeter (about 6% error). Their observation is that F4 is more reliable as a source of information about VTL than the lower formants. [See also Reby & McComb's, 2003, method of finding vocal tract length from formants.] Pisanski et al. (2014) found that VTL estimates are not particularly well correlated with height (r ~= 0.3), though listeners do have quite consistent perceptual judgements about talker height and these judgements are better correlated with vocal tract length than with weight (Smith & Patterson, 2005).

Clearly, there are phonetic details that distinguish speakers beyond just their gender and relative size of vocal tract. Among these, voice quality, the pattern of vocal fold vibration has received some attention (Ferrand, 2002; Harnsberger et al., 2010), as has a possible role of palate shape on vowel acoustics (Johnson, Ladefoged & Lindau, 1993), while acoustic differences due to nasal morphology (Guilherme et al, 2009; Subramaniam, et al. 1998; Maddux et al., 2017) seems like an unexplored area for fruitful research on talker differences.

Much of the research done on talker normalization has focussed on understanding how listeners must (it is assumed) map acoustic properties of speech produced by men and women to 'talker-independent' linguistic representations. Within-gender talker variation has not been a point of concern for most theorists, despite the fact that men do differ from other men, and women differ from other women on in terms of vocal tract and vocal folds; and also in terms of individual differences in dentition, palate, voice, and nasal cavity that are not relatable to gender.

3 The vowel normalization problem

In Peterson and Barney's (1952) seminal study of American English vowels (Figure 1, upper left panel) there is an impressive degree of overlap in the locations of the vowels in the two-dimensional F1/F2 "vowel space". And so we have a problem to explain: How can listeners correctly identify vowels which overlap so much in the vowel space?

It turns out that in a statistical sense, the answer to this question is to add more acoustic information about the vowel than just the F1 and F2 frequencies (Hillenbrand et al., 1995). With dynamic information of F0 together with F1-F4, identification of carefully produced vowels is as good as listeners. The theoretical challenge, then, is to understand the neuro-cognitive mechanisms that make use of these complex and seemingly incomparable acoustic patterns.

Consideration of a practical problem in describing the vowel systems of languages and dialects will set the stage for our discussion of perception. The practical problem is that in describing the vowels of a language or dialect we feel that we must 'normalize' some of the differences between talkers so that a more general shared talker-independent linguistic pattern of vowel production can be seen. This is a part of the theoretical discussion too, because children learn to "imitate" the speech of their speech community, despite the fact that they produce speech that is acoustically and auditorily very different from the adults in the community. In this section we will review some practical vowel normalization methods, and in the sections after this one we will discuss the perceptual processes that listeners may use to accomplish speech recognition in the face of talker variation.

As mentioned above, much of the variation between talkers in vowel acoustics is due to the differences in vocal tract length between men, women, and children, and it is possible to use acoustic measurements to estimate the talker's vocal tract length. We can then normalize vowel formant measurements relative to vocal tract length, thus removing one of the main sources of the acoustic difference between talkers.

Nordstrom & Lindblom (1975) used the frequency of F3 in open vowels (where F1 is greater than 600 Hz) to estimate the speaker's vocal tract length and from this scaled all speakers onto a vocal tract of a "standard" length (i.e. male). Wakita (1977) also used higher formants exclusively to estimate vocal tract length. Lammert & Narayanan (2015) found support for Nordström & Lindblom's intuition that higher formants may be more reliable indicators of vocal tract length, compared with F1 and F2. However, they found that methods relying only on F3 and F4 were not as accurate as models using the first four formants. Reby & McComb (2013) also used all of the available vowel formant measurements in their approach to measuring vocal tract length, calculating ΔF, the average interval between formants as the slope of a line relating frequency with formant number (see also, Fitch, 1997). In the framework of Lammert & Narayanan (2014), we can implement this with (1) and then calculate vocal tract length by (2).

1) $\Delta F = 1/mn \sum_{j}^{m} \sum_{i}^{n} [F_{ij}/(i-0.5)]$, where i = formant number and j is token number.

2) $VTL = c/2\Delta F$, c = 34000 cm/s

Vocal tract length normalization (Nordström & Lindblom, 1975) is a uniform scaling method. This means that speech is mapped onto a reference vocal tract by estimating the length of the speaker's vocal tract and then scaling the formant frequencies as if they had been produced by the reference tract with a single scale factor based on vocal tract length. Alternatively, the ΔF approach simply expresses vowel formant frequencies in terms of an acoustic measure of vocal tract length - the average interval between formants. This also is a uniform scaling technique with a single scale factor for all of the vowels and formants produced by the speaker.

Fant (1975) observed that uniform vocal tract length scaling may not be quite right because male and female vowel spaces can be made to match better if there are many scale factors (one for each formant of each vowel). His proposed non-uniform scaling method allowed for vowel-specific and formant-specific scaling factors. One motivation for non-uniform scaling in Fant's view is that, in addition to vocal tract length differences, men and women differ in the relative lengths of their oral and pharyngeal cavities. So the same constriction locations may lead to different formant patterns depending on the vowel. It is interesting, though, that in Fant's (1975) data, the scaling factors for different vowel/formant combinations are correlated with the formant frequency. This suggests a non-linearity in talker differences that might be captured by using a non-linear, auditory frequency scale.

Figure 1 shows the vowel spaces that are obtained when various vowel normalization algorithms are applied to the Peterson & Barney (1952) vowel data [many thanks to Santiago Barreda for the collection of datasets and analysis tools that he has made available in the R package *phonTools*]. There is a proliferation of normalization algorithms (and we have added one, based on Reby & McComb's (2013) ΔF method of finding vocal tract length) and it is apparent from the figure that many of the algorithms produce substantially similar results, so some discussion of their shared assumptions is worthwhile.

For each vowel normalization algorithm we ask three questions. (1) Is the information used to scale the vowel formants intrinsic to the vowel, or must we gather extrinsic information about the speaker from context in order to calculate the normalized values? (2) Are separate scaling factors used for the different formants? And (3), what frequency scale is used in making the calculations?

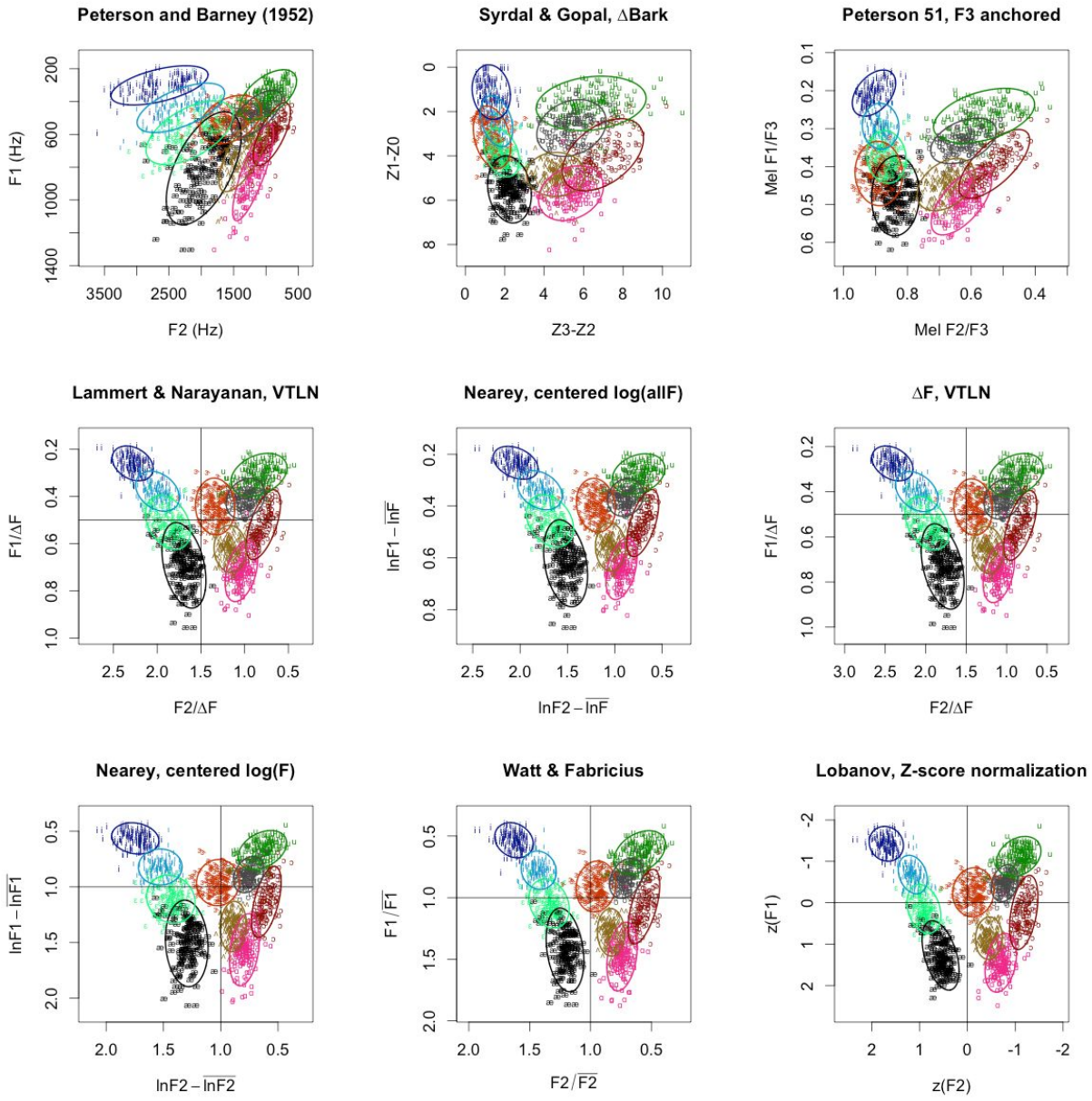

Figure 1. Vowel formant measurements from Peterson and Barney's classic 1952 study, and eight different ways to normalize them.

Most of the algorithms illustrated in figure 1 use vowel extrinsic information to perform vowel normalization. In fact, the only ones that don't are Syrdal & Gopal's (1986) auditory formant distances method, and Peterson's (1951) F3 anchoring method. The Vocal Tract Length Normalization (VTLN) procedures by Nordström and Lindblom (1975), Lammert & Narayanan (2015), and the ΔF method use information that is extrinsic to the token being scaled. The only difference between these is in how vocal tract length is found, where Nordström and Lindblom estimated vocal tract length from F3 in vowels with F1 higher than 600Hz, the more recent methods use all of the formant measurements taken from a speaker to

estimate the VTL. The point is, data from many tokens produced by the speaker are used to calculate the scale factor.

Similarly, the methods that use statistics calculated over all tokens from a talker to then normalize that talker's tokens are 'extrinsic' vowel normalization methods. This includes algorithms that use the mean values of formants (Lobanov, 1971; Fabricius, et al., 2009; Nearey, 1978). Gerstman's (1968) range normalization method falls in this category as well. Miller's (1989) formant ratio model is an unusual hybrid because he uses an extrinsic measure of voice pitch (which he called the sensory reference - SR) calculated over a span of prior speech, while the remaining parameters are local to the vowel being classified. For a database of isolated word reading such as the Peterson & Barney (1952) set this procedure is not very necessary, but in running speech F0 variation should be leveled out somehow (Johnson, 1990a) and the SR is intended to do this.

The middle row in Figure 1 shows methods that use uniform scaling. A single scale factor is applied equally to both F1 and F2. In vocal tract length normalization (VTLN) this factor is related to the calculated length of the vocal tract. The Nearey uniform scaling method that is shown here uses the geometric mean of the log formants (F1, F2, & F3) as the scaling factor, so the interpretation of the units on the axis is not very straight-forward.

The bottom row in Figure 1 shows methods of vowel normalization that use "non-uniform" scale factors (Fant, 1975). In each of these, a scale factor is calculated separately for each formant. In fact, the scale factor in each is related to the mean value of the formant being scaled (which is marked with the cross-hairs in the graph). In z-score normalization and in Watt & Fabricius' method the center is calculated in Hz, while in Nearey's (1978) approach the center is calculated in log(Hz). The unit of measure in these methods is also different, being the standard deviation of the formant in z-score normalization and the ratio between the formant and the mean formant in the other methods.

Several researchers (Hindle, 1978; Disner, 1980; Adank et al., 2004) have compared vowel normalization methods using classification accuracy of vowels and reduction of speaker information as criteria for evaluating the different methods. As seems obvious in Figure 1, there is a great deal of similarity between many of the normalization methods, and a set of support vector machine (SVM) classification models confirms their similarity. The top portion of table 1 shows identification performance by algorithms that use two dimensions, centering and scaling F1 and F2 in various ways. The bottom half of the table compares algorithms that use three dimensions.

It is noteworthy that the VTLN methods differ quite substantially from each other in vowel classification accuracy even though they use the same normalization method. This highlights the fact that an accurate estimate of vocal tract length is crucial in this approach, as in the others. It is also noteworthy that none of these normalization methods completely removes talker-type information. Classification by type (e.g. man, woman, child) is always better than chance.

Table 1. Percent correct identification by support vector machine (SVM) models of Peterson and Barney (1952; PB52) and Hillenbrand et al. (1995; H95) vowels and talker types (man, woman, child [MWC], or man, woman, boy, girl [MWBG]) by different vowel normalization methods. The best algorithms will maximize vowel classification and eliminate speaker information in the representation. If there is no speaker "type" information in the representation, then the % correct for MWC in PB52 will be less than 40% and for MWBG in H95 will be less than 31% (estimated by a permutation test).

Method	Type	PB52 % corr (vowels)	H95 % corr (vowels)	PB % corr (MWC)	H95 % corr (MWBG)
No normalization (F1 and F2)	NONE	77.3	62.9	66.7	53.2
Mean λ (Patterson & Irino)	Intrinsic	79.5	67.0	49.9	44.9
F3 anchor (Peterson 51)	Intrinsic	78.6	71.3	52.7	44.7
F1 anchor (Peterson 61)	Intrinsic	79.4	72.0	49.7	40.9
Mean F* anchor (Sussman)	Intrinsic	80.1	72.3	49.5	41.3
VTLN (Nordström & Lindblom)	Uniform	82.5	72.7	49.8	41.9
VTLN (Lammert & Narayanan)	Uniform	87.6	77.5	51.1	43.0
Mean F*, log difference (Neary 2)	Uniform	88.0	77.8	51.7	42.8
Range Normalization (Gerstman)	NonUniform	85.2	74.8	47.6	40.6
VTLN (Delta F)	Uniform	88.2	78.1	50.9	42.9
Mean F, log difference (Nearey 1)	NonUniform	90.9	80.1	51.6	42.4
Mean F, ratio (Watt & Fabricius)	NonUniform	90.8	80.7	50.8	41.4
Z-score normalization (Lobanov)	NonUniform	92.6	84.4	49.3	39.8
No normalization (F1, F2, F3)	NONE	86.5	76.9	83.4	69.9
Mel scale (F1, F2, F3)	NONE	86.4	76.8	84.3	70.3
Mean λ (Patterson & Irino)	Intrinsic	82.0	72.7	67.7	55.5
Formant ratios (Miller)	Extrinsic	86.0	78.1	59.2	52.8
Bark differences (Syrdal & Gopal)	Intrinsic	83.9	77.1	58.3	44.0
No normalization (f0, F1, F2, F3)	NONE	90.0	81.3	94.8	75.5

The Watt & Fabricius method (Fabricius et al., 2009) was designed to be quite careful about the selection of the midpoint of the talker's vowel space by selecting a subset of vowels that will sample the extremes of the available formant space for each speaker. It wasn't clear that this care in selecting the midpoint of the vowel space is actually necessary (see e.g. Bigham, 2008), so we estimated the center of the space as the mean values of F1 and F2 (as is done in z-score normalization and in the Nearey 1 method). The non-uniform extrinsic methods essentially estimate the vocal tract length separately with each formant -- an estimate of VTL from F1, and a separate one from F2, etc. For the purposes of this chapter it is important to realize that extrinsic vowel normalization algorithms succeed by shifting talkers onto the same center, and scaling their range of variation in some way. This is a logical description of the talker normalization process, the first of Marr's (1982) levels of analysis.

Table 1 also shows classification accuracy without any vowel normalization when three (F1, F2, F3) or four (f0, F1, F2, F3) acoustic vowel measurements are provided. The result largely agree with those reported by Hillenbrand et al. (1995), who found that a linear discriminant function using only F1 and F2 could correctly classify only 68% of American English vowel tokens. But, when F3 was added to the function, the correct classification jumped to 81%, and 79% were correctly classified when the set of predictors was f0, F1 and F2. They found that a discriminant function with f0, F1, F2, and F3, taken from three temporal locations in each vowel, plus the vowel duration could correctly classify 95% of the tokens. This indicates that each vowel contains within itself information that, when appropriately utilized, can correctly identify the vowel without any vowel extrinsic information. Table 1 also shows that when several raw dimensions are used to represent vowels, the representation supports both talker and vowel classification.

So, we have these two observations about the information conveyed by vowels. (1) Extrinsic information can be used to place the main cues for vowel identity into a vocal tract length frame of reference that facilitates vowel classification on the basis of just two or three acoustic attributes. (2) But at the same time, information that is intrinsic to the token, the f0 and higher formants, may also provide such a frame of reference for classification. We turn now to consider how listeners seem to make use of these complementary sources of information in speech perception.

4 Intrinsic normalization

4.1 Vowels as Formant Patterns

Formant ratio theory was proposed by Lloyd (1890). He said, "There is no way in which single isolated resonances can be imagined strongly to differ except in absolute pitch. But when it has been shown that the principal vowels all probably possess *two* resonances we are at once delivered from the necessity of any such inference. It at once becomes conceivable that the fundamental cause of any given vowel quality is the relation in pitch between two resonances, irrespective of any narrow limit in absolute pitch." (Lloyd, 1890, p. 169).

This idea has been echoed many times in subsequent studies. For example, Potter & Steinberg (1950) stated that in vowel perception "a certain spatial pattern of stimulation on the basilar membrane may be identified as a given sound regardless of position along the

membrane." They compared vowel perception to the perception of musical chords saying, "the ear can identify a chord as a major triad, irrespective of its pitch position." (p. 812)

Traunmüller (1981, 1984) also concluded that "perception of phonetic quality" can be "seen as a process of tonotopic Gestalt recognition" (1984; p. 49). Bladon, Henton & Pickering (1986) implemented a whole-spectrum matching model of vowel perception of this idea by simply sliding auditory spectra up or down on the frequency scale. This is similar to the procedure used for vocal tract length normalization (VTLN) in automatic speech recognition (ASR) (e.g., Garau et al., 2005; Kinnunen & Li, 2010), with an important difference. The spectral representation most commonly used in ASR - the Mel-Frequency Cepstrum - does not include f0.

In line with Fant's (1975, p. 16) observation that "uniform expansions or contractions of the tract leave resonance-frequency ratios intact", and harking back to Lloyd's (1890) conception, Sussman (1986) and Miller (1989) proposed to use formant ratios as dimensions in their models of vowel perception. Syrdal and Gopal's (1986) use of formant differences on an auditory scale is comparable to this - as Miller (1989) pointed out, log(F2/F1) is equivalent to log(F2) - log(F1). However, unlike Peterson's (1951, 1961) and Sussman's (1986) conception of formant ratio theory, these models, along with Traunmüller and Bladon et al., incorporate the fundamental frequency (f0) in their formant ratio representations. This makes sense considering how prominently the harmonics of the fundamental figure in auditory spectra.

Sussman (1986) proposed a specific neuronal type of process to instantiate a formant ratio theory of vowel normalization. In his approach, which was inspired by studies of bat echo location (Suga et al., 1983), each formant is encoded relative to a sum of the formants in a simple neural circuit.

4.2 F0 normalization

Evidence for the perceptual effects of intrinsic information in vowels comes from studies of both the perceptual effects of f0 (the fundamental frequency of voicing) and of higher vowel formants.

The perceptual effect of vowel f0 was studied by Miller (1953) who found that the perceptual category boundary between vowels shifted when the f0 was doubled from 120Hz to 240Hz. Fujisaki and Kawashima (1968) studied this further and found F1 boundary shifts of 100Hz to 200Hz when f0 was shifted by 200 Hz. Slawson (1968) estimated that an octave change in f0 (doubling) produced a perceived change in F1 and F2 of about 10-12%. The direction of these effects mirrors the correlations found in speech production, namely that as f0 increases the perceived values of the formants also increase.

Barreda & Nearey (2012) found that f0 impacts vowel perception indirectly through the "perceived identity of the speaker" (Johnson, 1990a), rather than directly as envisioned by Syrdal & Gopal (1986) or Miller (1989) who included a term for F1/f0 ratio in their representations of vowels. This indirect effect is mirrored in the neural representation of speech as we will see below.

4.3 Higher formant normalization

It has also been reported that the boundaries between vowel categories are sensitive to the frequencies of a vowel's higher formants (F3-F5). Fujisaki and Kawashima (1968) demonstrated an F3 effect with 2 different vowel continua. An F3 shift of 1500 Hz produced a vowel category boundary shift of 200 Hz in the F1-F2 space for a /u/-/e/ continuum, but a boundary shift of only 50 Hz in an /o/-/a/ continuum. Slawson (1968) found very small effects of shifting F3 in six different vowel continua. Nearey (1989) found a small shift in the mid-point of the /ʊ/ vowel region when the frequencies of F3-F5 were raised by 30%, but this effect only occurred for one of the two sets of stimuli tested.

A possible explanation for some of the inconsistencies found in this literature was offered by Johnson (1989) who also found an F3 boundary shift, and attributed it to spectral integration of F2 and F3 (Chistovich,1979) because the F3 frequency manipulation only influenced the perception of front vowels (when F2 and F3 are within 3 Bark of each other) and not back vowels which have a larger frequency separation of F2 and F3. Note that this finding suggests that an aspect of general auditory processing (spectral smearing) impacts perceptual normalization, perhaps circumventing the need for an exemplar-based account of talker normalization even for isolated vowels.

4.4 Neural correlates of intrinsic normalization

When it comes to the neural infrastructure that is involved in the processing and/or representation of speech sounds, generally, three questions are relevant: 1) What is the dominant representation of vowels? Abstracted phonemes, or rather their underlying acoustic/phonetic properties? 2) How veridical are these representations? Are their acoustic properties preserved, or are they normalized? 3) If normalized representations exist, what properties of auditory cortex processing allows them to emerge? As will become clear, the first question can be best answered, as the dominant representations of speech sounds in auditory cortex seem to clearly reflect acoustic/phonetic representations. The second question can be partially answered. It appears that the brain does give rise to warped vowel-identity related representations in a separate processing stream than the one for speaker-identity related representations, although it remains unclear whether these representations become completely or only partially separated. For the third question, only suggestive evidence can be offered. These findings will, however, be useful in guiding future research into this topic.

4.4.1 The basic auditory processing hierarchy

To understand how speech sounds are processed by the human brain, it seems useful to first briefly discuss some of the most relevant properties of early auditory processing, focussing especially on representations in auditory cortex since that is where more complex and speech specific representations become dominant. Figure 2 provides a visualization of the anatomical cortical landmarks of the regions involved. The majority of auditory information reaches the cortex through ascending connections to Primary Auditory Cortex (PAC) which is mostly situated within the Sylvian fissure on so-called Heschl's gyrus. A dominant property of PAC is that it partly inherits the tonotopic organization of the cochlea (a place coding of acoustic frequencies). That is, adjacent cortical areas on PAC are sensitive to slightly different

frequencies in the auditory signal (Bauman et al., 2013; Bitterman et al. 2008; Humphries et al., 2010; Formisano et al., 2003; Moerel et al. 2012; Saenz & Langers, 2014). Many of the acoustic cues that are critical for the perception of speech, such as formants, formant transitions and amplitude modulations, are represented on PAC as spatial and spatiotemporal patterns of activation (e.g., Young, 2008). The most important cortical structure receiving direct information from PAC is the broader secondary auditory cortex. This includes the planum temporale and planum polare, both situated within the Sylvian fissure, and the laterally exposed Superior Temporal Gyrus [STG]. Patches of tissue in secondary auditory cortex are often described as behaving like filters that are sensitive to increasingly complex spectro-temporal information (i.e., combinations of frequencies and/or frequency sweeps), like that observed in natural speech. It is generally observed that the processing of sound becomes increasingly speech-specific for patches of tissue located further away from PAC (see Obleser & Eisner, 2008; Price, 2012, Liebenthal et al., 2014; Turkeltaub & Coslett, 2010; DeWitt & Rauschecker, 2012; Overath, McDermott, Zarate, & Poeppel, 2015), especially in ventral and anterior directions on the STG.

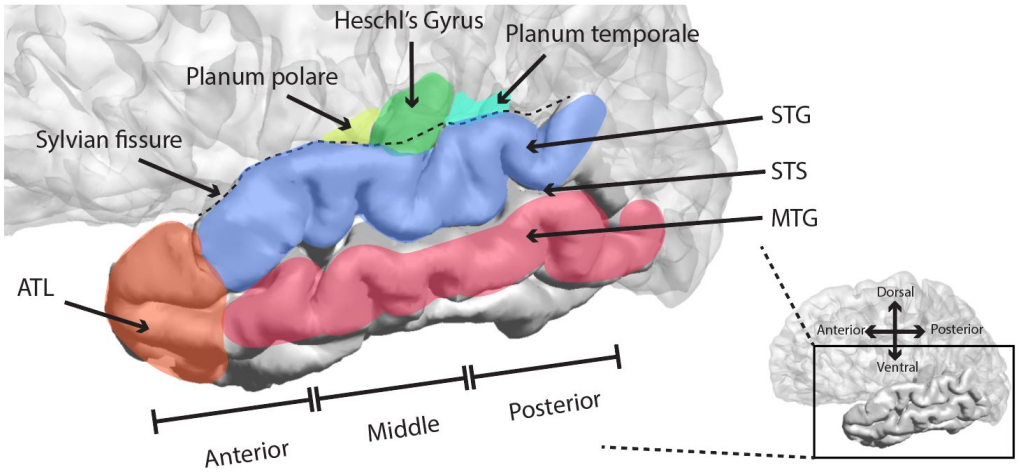

Figure 2. Anatomical landmarks of the temporal lobe on and around the regions involved in early-speech sound processing. Regions outside the Temporal Lobe are displayed as transparent, allowing for the visualization of Heschl's Gyrus and Planum Polare and Planum Temporale which are all situated inside the Sylvian fissure. Abbreviations: Superior Temporal Gyrus (STG); Superior Temporal Sulcus (STS); Middle Temporal Gyrus (MTG); Anterior Temporal Lobe (ATL).

A dominant idea has been that beyond PAC processing there is a stream of activation that is especially involved in the transition from acoustic-phonetic sound representations towards the activation of lexico-semantic representations, resulting in comprehension. This flow of information involves the spreading of activation from PAC and closely surrounding regions towards anterior (ATL) and ventral regions (MTG) of the temporal lobe, which are often thought to be involved in lexical level processing (Davis & Johnsrude, 2003; DeWitt & Rauschecker, 2012; Hickok & Poeppel, 2007; Scott, Blank, Rosen, & Wise, 2000). This flow of information is typically contrasted with a second flow of information directed outside of the temporal lobe that is thought to be involved in sensory–motor integration and phonological working memory (the so-called dorsal stream). The functional properties of this second flow of information will not be

discussed in further detail detail here (see Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Rauschecker & Tian, 2000; Scott & Johnsrude, 2003 for discussion).

As suggested earlier, one of the important questions is what the dominant representational form of speech sounds is. This could involve a representation based on syllables, phonemes, gestures or acoustic-phonetic features, to name a few. In a recent investigation, Mesgarani et al., (2014) observed that focal patterns of cortical activity displayed selectivity for phonetic features. Activity recorded from individual electrodes would display a reliable response to a set of phonemes (e.g. plosives /d/, /b/, /g/, /p/, /k/, and /t/, or sibilants /ʃ/, /z/ and /s/, or low-back vowels /a/ and /aʊ/, or high-front vowels and glides (/i/ and /j/). None of the electrodes from which they recorded displayed a preference for single phonemes. These observations suggest that at the level of the STG the human auditory cortex represents speech sounds as organized by acoustic-phonetic features (see also Arsenault & Buchsbaum, 2015; Steinschneider et al., 2011). These findings align with earlier single neuron recordings (Chan et al., 2014; Creutzfeldt et al., 1989).

4.4.2 The cortical separation of vowel-types and voice properties

Based on recordings of these spatially distributed acoustic-phonetic feature representations vowels can be separated. However, separability is thus closely related to acoustic-phonetic similarity. Indeed, fMRI research has demonstrated that classification techniques that are sensitive to spatially distributed patterns of activation (e.g., multi voxel pattern analysis) allow for a quite accurate separation of at least the corner vowels /i/, /a/, and /u/ in bilateral auditory cortex (e.g., Formisano, 2008). Furthermore, in MEG experiments it has also been observed that those vowels that are acoustically most distinct, also give rise to most dissimilar responses (e.g., Obleser et al., 2003; Shestakova et al. 2004). Mesgarani et al. (2014) further confirm this pattern as they observed a high correlation ($r = 0.88$) between acoustic distances in the F1-F2 space of pairs of vowels and their resulting spatial neural activity differences.

In addition to the observation that vowel-types can be distinguished based on their neural responses it has also been demonstrated that the neural representation of vowel-types and the representation of speaker-specific information (presumably related to F0 and higher formants) involve partially non-overlapping cortical patches (e.g., Formisano et al. 2008; Edmonds et al., 2010; Bonte et al., 2014). For example, Edmonds et al., 2010, presented listeners with steady vowel (or vowel-like noise) sounds that transitioned into other vowel sounds and/or sounds cueing a speaker change (with no gap in between items), while recording EEG. These authors modified formant frequencies to induce the percept of the vowels /a/, /e/, /i/, /o/, /u/, while also modifying the overall ratio between the different formants. They reported larger ERP deflections, estimated to originate from PT and PP, when one vowel transitioned into another vowel (but spoken by the same speaker) than when a specific vowel transitioned into the same vowel but spoken by another speaker. Similar observations have recently been reported based on MEG (Andermann et al., 2017). Those authors reported spatially separable sources that related to changes in pitch (estimated to originate from Heschl's Gyrus) and vowel type and mean formant frequency (both localized to PT). A related result has been reported in an MEG study by Monahan and Idsardi (2010). They relied on the observation that the response

latency of the M100 component tends to decrease (i.e, arise earlier) when the frequency of F1 is nearer 1000 Hz. They observed that latency of this component was not only dependent on F1, but was also sensitive to the frequency of F3 (for two out of three vowels tested). They concluded that these results provide evidence that the M100 is actually coding the F3/F1 ratio - a parameter in Peterson's (1951, 1961) vowel normalization method (see table 1 above). These studies this support the notion that different dimensions of the speech signal may become separately represented, and that those representations are partly "invariant" with respect to changes on the other dimensions. Other support for this notion comes from the observation that the robustness of these separate representations is modulated by task demands. fMRI research has shown that more accurate cortical vowel classification is observed in "vowel identification" tasks, and better cortical speaker identity classification in "speaker identification" tasks (Bonte et al., 2014; von Kriegstein et al., 2010). Moreover, it has been shown that models trained on the BOLD data from listening to one set of speakers can be used to accurately classify vowel identity from held-out speakers (Formisano et al, 2008).

It bears mentioning, however, that these studies do not seem to provide unequivocal evidence for normalized vowel representations. That is, while the separation of vowel-type and speaker-identifying information is promising it is important to appreciate that these properties are also partly represented as different parts of the speech signal (F1-F2 for vowel type and F0; and higher formants for speaker information). Hence, it remains unclear whether the separability of these properties in cortical activation simply reflects the specific sensitivities of different patches of cortex, some responding preferentially to lower formant frequency ranges and others responding to pitch and/or higher formant frequencies. Moreover, a number of these findings reported above appear somewhat inconclusive. For example, the F3/F1 ratio interpretation offered by Monahan and Idsardi (2010) required a significant theoretical hedge to explain why [ɒ] and [ɛ] showed the effect while [ə] did not. In the study of Edmonds et al., 2010, the overall acoustic changes in F1 and F2 seem to have been larger for the most extreme vowel changes than for the most extreme speaker changes, which could explain the larger ERP deflections observed for vowel-type changes. More generally, it is also clear that while perhaps *partly* non-overlapping, the representation of voice acoustics and vowel-type information in fact seem to involve mostly the same general cortical regions (e.g. Chandrasekaran et al., 2011). In Formisano et al. (2008) Figure 4, and Bonte et al. (2014) Figure 5, for example, it is clear that the vowel classification maps do still retain speaker gender information.

Some of the uncertainty in interpreting the studies discussed above arises from the fact that those studies were partly restricted in either spatial or temporal resolution, which presents a challenge to being able to track the representation of specific speech cues. A recent ECoG study has addressed intrinsic talker normalization in the perception of intonation contours (Tang et al., 2017). Tang et al presented listeners with sentences that were each synthesized to contain four different intonation contours. These sentences were synthesized with male and female voices, thereby creating large absolute differences in pitch frequencies. Tang et al. (2017) found that auditory cortex contains cortical patches of tissue that follow the speaker-normalized pitch contours. That is, they responded in the same way to a linguistically identical pitch contours, irrespective of absolute F0 and phonetic content. Moreover, only very few electrodes demonstrated sensitivity to absolute, instead of normalized pitch. This study thus

demonstrates what the representational outcome of a successful normalization process for vowels *could* look like at a more fine-grained level of representation. More importantly, it confirmed that auditory cortex indeed generates normalized representations for at least some speech cues, buttressing the findings of vowel-type normalization reported above.

4.4.3. How may frequency-independent coding of formants emerge in auditory cortex?

As discussed earlier, a number of normalization approaches have relied on a relative coding scheme. That is, a scheme where formants are interpreted not as absolute features but rather on the basis of their relation to each other and perhaps to the fundamental frequency. Assuming that these normalized representations indeed exist, an important next question is what properties of auditory cortex processing may give rise to them. Although this is still a largely open question, in the following we will discuss some properties of auditory cortex processing that could play an important role in achieving such a format.

The speech signal contains both fluctuations in the overall amplitude envelope (the spacing and prominence of peaks in the amplitude envelope) and fluctuations in the spectral envelope (i.e., the spacing and prominence of peaks in the power spectrum). The ΔF measure of formant spacing that was discussed above as a computational vowel normalization factor (Table 1) is related to this notion. Research on auditory processing in animals has demonstrated that patches of tissue in human auditory cortex display tuning for specific combinations of spectral and temporal modulation frequencies (e.g., Depireux et al., 2001; Wooley et al., 2005; Nagel & Doupe, 2008). More recently, it has been demonstrated that this is also a dominant property of human auditory processing (Hullet et al., 2016). Those authors demonstrated that the human STG broadly displays an anterior to posterior organization of different spectro-temporal tuning profiles. Posterior STG sites displayed a preference for speech sounds that have relatively constant energy across the frequency range (low spectral modulation), but which are temporally changing at a fast rate. Anterior STG sites displayed preferences for speech sound sequences that show a high degree of spectral variation across the frequency range (high spectral modulation) but which are temporally changing at a slow rate. In support of this finding, BOLD response patterns of auditory cortex can be quite accurately predicted based on models that consider a combination of spectral properties, spectral modulations and temporal modulations in the acoustic signal (Santoro et al., 2014; 2017). This observation may be important to normalization because modulation frequencies are independent of specific frequencies, and rather represent the *pattern* of peaks and troughs across a range of frequencies. Hence, this property of auditory cortex processing has the potential to play a role in the frequency-independent representation of formant patterns. Relatedly, it has often been observed that vowels that display a large separation between the first and second formant result in overall larger N1m responses (Diesch and Luce, 1997, 2000; Shestakova et al. 2004). It has been suggested that this observation can be partly explained by formant inhibition, i.e., that closely neighboring formant peaks engage in reciprocal inhibition of responses, resulting in overall weaker cortical responses (Diesch and Luce, 2000; Obleser et al., 2003; Ohl & Scheich, 1997). This observation may be closely related to spectro-temporal modulation tuning, because differences in the distance between formants are also reflected as different spectral modulations.

5 Extrinsic normalization

Our discussion so far has focussed on how the perceptual system can compensate for talker variability by integrating different, co-occurring, auditory cues in the speech signal. However, speech sounds rarely occur in isolation. Typically we hear speech sounds in the context of some preceding and following speech sequences, and perhaps we are looking at the person we are talking to as well. This is important because such context can provide constraints on the possible interpretations of speech cues. That is, "knowing" that you are listening to a tall male speaker may enhance one's expectation of that speaker's' formant and pitch ranges (e.g. Joos, 1948). Indeed, a considerable literature has demonstrated that listeners use such acoustic and visual contextual information when interpreting speech sounds.

5.1 Extrinsic vowel normalization

The first demonstration of the fact that speech sound perception is highly dependent on acoustic properties of preceding context came from a series of experiments by Ladefoged and Broadbent (Broadbent, Ladefoged & Lawrence, 1956; Ladefoged & Broadbent, 1957). Their participants were asked to listen to synthesized versions of the words "bit" "bet" "bat" and "but". These target stimuli were presented in isolation or were preceded by a precursor phrase in which the speaker voice properties had been altered (by shifting the overall formant range to higher or lower frequencies). Quite strikingly, it was observed that vowel perception was strongly dependent on the voice properties in a preceding sentence. A target vowel that had been predominantly identified as "bet" when presented in isolation, was overwhelmingly identified as "bit" (which has a lower F1 than "bet") when the preceding sentence had a high F1 range. Listeners thus seemed to have adjusted the expected dynamic formant range when interpreting the incoming target vowels. Such normalization to a particular speakers' voice properties has since been replicated on various occasions, demonstrating that it generalizes across languages (Sjerps & Smiljanic, 2013), and that it applies to the perception of different spectral cues such as F1 and F2 (Nearey, 1989; Ladefoged, 1989, Sjerps & Smiljanic, 2013; Darwin et al., 1989; Watkins and Makin, 1994, Mitterer, 2006; Reinisch & Sjerps, 2013), but also to F0 (in the context of lexical tone: Moore & Jongman, 1997; Cantonese: Wong & Diehl, 2003; Leather, 1983; Fox & Qi, 1990; Jongman & Moore, 2000; Peng et al., 2012; Zhang et al., 2012; 2013; Francis et al., 2006; Lee et al., 2009), spectral tilt (Kiefte & Kluender, 2008), and duration cues (see e.g., Miller, 1984a; 1984b; Miller & Grosjean, 1981; Kidd, 1989; Reinisch, et al. 2011a; 2011b; Summerfield, 1981; Newman & Sawusch, 2009; Sawusch & Newman, 2000; see Miller & Liberman, 1979; Toscano & McMurray, 2015; Dilley & Pitt, 2010; Morrill et al., 2014; Pitt, Szostak, & Dilley, 2016). These results have given rise to a broad range of studies attempting to better understand what properties of perception give rise to these influences. Given the scope of this chapter, we will focus here only on contextual influences on vowel formants, assuming that normalization of other cues involves different functional mechanisms. Within the domain of extrinsic normalization of formants, two types of influences have been established: 1) Low-level auditory processes that enhance perceptual contrast between acoustic context and a target

sound, and, 2) Higher-level influences that depend on acquired knowledge about the relation between talker properties (such as gender) and the acoustic realization of speech sounds.

5.2 Mechanisms of extrinsic normalization.

5.2.1 *A role for auditory contrast*

In a classic study of how context-acoustics affect subsequent perception, Watkins and Makin (1994) carried out a very similar experiment to that of Ladefoged and Broadbent (1957), except that instead of shifting the formants of the context sentences they filtered a context sentence such that its Long-Term Average Spectrum (LTAS) matched that of either a low- or high-frequency average F1 carrier sentences (but without directly altering formant center frequencies themselves, see also Watkins, 1988). They observed qualitatively similar shifts in category boundaries as those observed by Ladefoged & Broadbent. Moreover, Watkins (1991) applied similar filters to a speech-shaped noise signal and used those as preceding contexts and observed similar effects as well (although numerically smaller, see below). Watkins and Makin (1994) thus argued that it was not the range of the context's F1 frequency that shifted target perception, but rather its LTAS, suggesting that the influence of context acoustics on vowel category perception could better be explained by an "inverse-filtering heuristic": Reliable spectral properties of a precursor, as reflected in its long-term spectral characteristics, are filtered out of the target sound before relevant acoustic properties are extracted for further processing.

Importantly, Watkins and Makin (1994) also argued that LTAS-based normalization may in fact result from contrastive effects that originate from at least two separate auditory processing stages. The noise-carrier effects that Watkins (1991) observed were only found when the noise context and the subsequent target were presented to the same ear, and when there was only a small (≤ 160ms) silent gap between context signal and the target sound. With speech precursors, the influence of contexts *did* apply to contralateral presentation and also over longer silent intervals. Furthermore, some effects of perceptual streaming seemed to play a role as well. When speech contexts were presented with different interaural time delays than the targets (i.e, inducing differences in percept of location) then contrastive effects were reduced. Watkins thus suggested the existence of two stages in the auditory processing hierarchy that may induce LTAS based effects. A peripheral stage, perhaps similar to the type of "negative auditory afterimage" reported by Summerfield et al., 1984 (explaining unilateral noise effects), and a more central contrastive mechanism (explaining the contralateral effects obtained with speech). Only the latter, then, is argued to operate over longer time scales and may be more specific to speech or speech-like stimuli.

But even those higher-level (i.e., speech specific) normalization effects appear to operate on at least pre-lexical and potentially even general auditory levels of representation. Despite long silent intervals between context sentences and target sounds normalization is independent of listeners' familiarity with the context language (Sjerps & Smiljanic, 2013), and nonwords have also been found to induce normalization effects (Mitterer, 2006). Similarly, speech from one speaker can have a normalizing influence on speech from another speaker (Watkins, 1991). And, perhaps more strikingly, reversed speech sounds are as successful in inducing normalization as normal speech sounds (Watkins, 1991). Also, extrinsic normalization

effects are stronger for the portion of the vowel space where the contexts differ (Mitterer, 2006b), so that the influence of high vowels in context is restricted to high vowels in targets. It appears, then, that these effects can not be the result of learned associations between talkers and their phonetic realization or of "strategic" shifts in category boundaries. Indeed, extrinsic normalization effects have not only been observed in categorization designs, but also in discrimination tasks that do not require listeners to make category-judgements (Sjerps et al., 2013). Moreover, effects are independent of whether listeners attend to the contexts themselves (Sjerps et al., 2012; Bosker et al., 2017).

Generally, these auditorily driven normalization effects have been interpreted to support the notion that listeners are especially sensitive to acoustic change (Stilp et al., 2010; Kiefte & Kluender, 2008; Sjerps et al., 2011). That is, the auditory system may calibrate to reliable properties of a listening environment in ways that enhance sensitivity to less predictable (more informative) aspects of sounds (Alexander & Kluender, 2010). Indeed, normalization effects are contrastive in nature. That is, the typical pattern is that a *high* formant context sentence leads to an increase in the percept of a *low* formant target option, while a low formant context leads to more high formant target percepts.

5.2.2 Tuning in to talkers

Importantly, however, acoustic-contrast based effects cannot be the sole explanation of normalization effects. It has long been known that listeners use higher-level information when making judgements about speech sound categories (e.g., Evans & Iverson, 2004). Evans and Iverson asked participants to rate category goodness for vowels that were presented in the context of sentences spoken in two different regional accents (northern or southern English accent). They demonstrated that participants' perceived quality ratings depended on what they expected a speaker with a certain accent to produce. This effect cannot be explained by LTAS based effects, for example because this effect interacted with listeners' own dialectal background (Evans & Iverson, 2004). Moreover, Listeners also adjust category boundaries as a result of non-auditory information. Johnson, Strand and D'Imperio (1999) presented listeners with sounds from a synthetic "hood"-"hud" continuum (spoken by an androgynous voice). Participants who were told that they were listening to the speech of a female speaker had the vowel category boundary closer to the female speaker boundary than that obtained from listeners who were told that the were listening to the speech of a male speaker. Similarly, when these sounds were presented in combination with a male or a female picture, listeners responded with more talker-appropriate category boundaries. These findings suggest that perception is mediated by a representation of perceived talker identity (see also Walker et al., 1995; Schwippert and Benoit, 1997).

In addition to normalization approaches based on categorization, another method involves the presentation of word lists that are either spoken by the same speaker across a block, or spoken by different speakers on subsequent trials. The typical finding is that switching speakers results in lower identification accuracy (e.g., Verbrugge et al., 1976; Barreda, 2012; Magnuson and Nusbaum, 2007; Nusbaum and Morin, 1992). Moreover, lists that involve switching talkers result in overall longer reaction times (e.g., Choi et al., 2018) and larger talker-normalization boundary shifts (Johnson, 1990b). These results have led to the suggestion

that talker identity-based normalization involves a cognitively demanding process (see Barreda, 2012 for review).

5.3 Neural extrinsic talker normalization

An important difference between intrinsic normalization and extrinsic normalization is that the latter requires the system to achieve and maintain a stable representation of the speaker and its acoustic voice properties so as to provide a frame of reference for further interpretation. As outlined in the previous section, extrinsic normalization effects may arise as the result of at least two types of influences: auditory-driven contrastive processes and higher-level influences of expected speech sound realizations based on known speaker properties. In the following we will discuss what is known about the cortical processing of these properties, and whether they may affect the cortical representation of speech sounds.

5.3.1 The representation of talker voice properties talker identities in cortical processing

As suggested earlier, the acoustically-driven normalization effects have often been interpreted in the framework of contrast enhancement. One way in which such contrast enhancement may be implemented is through neural processing properties such as adaptive gain control (Rabinowitz et al., 2011) or Stimulus Specific Adaptation (Ulanovsky et al., 2004; Pérez-González et al., 2014). These mechanisms have typically been investigated in the context of adaptation to differences in loudness or the presence of background noise. However, they may play a role in extrinsic normalization. Specific neural populations that display tuning to the frequency of one context sentence (say, a high formant sentence in Ladefoged & Broadbent, 1957) may adapt when a listener hears that sentence. Such adaptation could then affect responsiveness of these populations during subsequent target processing which could bias cortical representations of subsequent target sounds away from the context-specific F1 range. Hence, this may provide a more mechanistic implementation of the “inverse filtering heuristic” suggested by Watkins & Makin (1994). Importantly, effects of stimulus specific adaptation and forward suppression become more dominant, and are longer lasting, at cortical levels of processing (Philips et al., 2017; Fitzpatrick et al., 1999). This could, in principle, explain why auditory-contrast based effects are typically stronger for speech sounds than for nonspeech sounds. Speech sounds induce considerably stronger cortical activation than nonspeech sounds (DeWitt & Rauschecker, 2012;Scott & Johnsrude, 2003; Price, 2012, for general review), which may increase the amount of shared neural infrastructure between target and precursor and hence also the amount of adaptation.

Apart from acoustically driven influences, a considerable amount of work has been devoted to investigating the cortical representation of talker-specific acoustic voice properties and talker identities. Indeed, patterns of activation in the temporal lobe allow for accurate dissociation between speakers. A number of studies, however, have suggested that there exists a separation between, one the one hand, regions that are involved in the immediate processing of the acoustics of a particular voice and, on the other hand, regions that are involved in the representation of talker identities (e.g. Andics et al., 2010; von Kriegstein et al., 2003; Meyers et al., 2017). The representation of token-specific voice acoustics has been found to involve bilateral temporal regions (e.g., STG & STS). These regions are, thus, in terms of topography,

largely overlapping with those regions involved in the representation of speech sounds more generally (although slightly more right dominated; e.g., Bonte et al., 2009; Bonte et al., 2014; Formisano, 2008; von Kriegstein et al., 2003). Though, von Kriegstein et al., (2006) also found that changes in vocal tract length are associated with changes in activity along the pre-cortical auditory pathway in the Medial Geniculate Body.

The representation of talker identities (or access to known voices), however, has most often been associated with activation in the right ATL (Andics, McQueen, & Petersson, 2013; Andics et al., 2010; Belin & Zatorre, 2003; Campanella & Belin, 2007), and processing in the IFG (e.g., Andics et al., 2013; Latinus et al., 2011, Pernet et al., 2015; Zaske et al., 2017). Especially the ATL seems to play an important, and heteromodal, role as lesions to that region are associated with a reduction in the ability to name famous faces and famous voices (Abel et al., 2015; Drane et al., 2013; Damasio et al., 1996; Waldron et al., 2014, see Blank et al., 2014 for review), indicating that auditory and visual based identity processing streams converge in the ATL. Myers & Theodore (2017) found that phonetic atypicality (a rather more aspirated /k/ than typical for English) provokes a heightened response in core phonetic processing areas (bilateral MTG and STG), while talker phonetic atypicality (a rather unusual /k/ than one has come to expect for a particular talker) is associated with deactivation in the right posterior MTG. Intriguingly, talker typicality also modulated connectivity between this deactivated MTG region and the left motor cortex.

5.3.2 The extrinsic rescaling of vowels in cortical processing

The existence of contrast enhancing properties along with the robust representation of talker identity information in ATL and IFG may thus allow for both auditory-contrast based and talker-identity based influences on vowel representations in auditory cortex. Although, as far as we aware, only a single study to date has investigated this, it does appear that speech sound representations in auditory cortex become normalized as a result of preceding context. Sjerps et al., (2011b) presented listeners with sequences of short (3-syllable) context-target pairs in an oddball EEG design. Target vowels involved "standard" sounds that were perceptually ambiguous between [ɛ] and [ɪ]: [$^{\varepsilon}_{\text{ɪ}}$]. The (infrequent) "deviants" consisted clear instances of [ɛ] and [ɪ] (an F1 distinction). Context bysillables (/papu/) were synthesized to contain either generally heightened or lowered F1 distributions. It was observed that after high F1 contexts, a shift from an ambiguous standard [$^{\varepsilon}_{\text{ɪ}}$] to clear deviant [ɛ] resulted in a larger neural mismatch signal than a shift to clear [ɪ]. The reverse pattern was observed when the context had a low F1 distribution. This pattern of results suggests that the context syllables induced a perceptual shift of the ambiguous standard, leading to smaller or larger mismatch signals (and lower or higher oddball detection scores). For example, the standard would sound more like [ɪ] (the low F1 option) after a high F1 context, hence reducing the mismatch with a deviant [ɪ] (and increasing the mismatch with [ɛ]). Importantly, the normalizing effect was observed as early as the N1 time window, which is more strongly related to the physical properties of the stimulus, than to participants' perceptual decisions (e.g., Roberts et al., 2004; Toscano et al., 2010; Näätänen & Winkler, 1999). This suggests that those effects arose early in cortical processing as an auditory contrast effect.

6. Conclusions

The picture that emerges in this review is that the process of talker normalization in speech perception is dispersed in several neuro-cognitive mechanisms. On the one hand, some basic properties of the auditory system in how sound is contextually coded may produce "normalization" effects. Another low-level phenomenon may involve the perception of the length of the talker's vocal tract, perhaps in terms of spectral modulation, or the number of spectral peaks in the bottom two-thirds of the auditory spectrum, which then is available to warp the auditory percept even precortically. The evolutionary need to code the size of con-specific individuals probably drove the emergence of vocal tract size perception long before the emergence of language.

And yet, the warping and filtering along the primary auditory stream is not the whole story. Behaviorally, we know that speech perception is also influenced by talker expectations running in parallel with the auditory stream (or even paradoxically in a somewhat separable stream within audition, as appears to be the case for the role of voice pitch). The field is relatively wide open for neural processing studies that explore the interaction of higher-level talker expectations and speech processing, and studies like Myers & Theodore's (2017) on how memory for specific talkers may interact, perhaps in multiple stages of processing, with the extraction of linguistic/phonetic information will reveal much about the neuro-cognitive mechanisms that support speech perception in a world filled with talker variation.

REFERENCES:

Abel, T. J., Rhone, A. E., Nourski, K. V., Kawasaki, H., Oya, H., Griffiths, T. D., ... & Tranel, D. (2015). Direct physiologic evidence of a heteromodal convergence region for proper naming in human left anterior temporal lobe. Journal of Neuroscience, 35(4), 1513-1520.

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *J. Acoust. Soc. Am.* **116**, 3099–3107.

Alexander, J. M., and Kluender, K. R. (2010). Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context. *J. Acoust. Soc. Am.* **128**, 3597-3613.

Andermann, M., Patterson, R. D., Vogt, C., Winterstetter, L., & Rupp, A. (2017). Neuromagnetic correlates of voice pitch, vowel type, and speaker size in auditory cortex. *NeuroImage*, 158, 79-89.

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

Arsenault, J.S. & Buchsbaum, B.R (2015) Distributed Neural Representations of Phonological Features during Speech Perception. *Journal of Neuroscience* **35**(2) 634-642; DOI: 10.1523/JNEUROSCI.2454-14.2015

Barreda, S. & Nearey, T.M. (2012) Direct and indirect roles of fundamental frequency in vowel perception. *J. Acoust. Soc. Am.* **131**(1), 466-477. DOI: 10.1121/1.3662068

Baumann, S., Petkov, C. I., & Griffiths, T. D. (2013). A unified framework for the organization of the primate auditory cortex. *Frontiers in systems neuroscience*, 7, 11.

Bigham, Douglas. (2008). Dialect contact and accommodation among emerging adults in a university setting. Ph.D. thesis, The University of Texas at Austin.

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., & Nelken, I. (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature*, 451(7175), 197-201.

Bladon, R.A., Henton, C.G. & Pickering, J.B. (1984) Towards an auditory theory of speaker normalization. *Language Communication 4,* 59-69.

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: merging evidence from patients and healthy individuals. Neuroscience & Biobehavioral Reviews, 47, 717-734.

Bonte, M.; Hausfeld, L.; Scharke, W.; Valente, G. & Formisano,E.(2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J. Neurosci.* **34**, 4548–4557. doi:10.1523/JNEUROSCI.4339-13.2014

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. Journal of Memory and Language, 94, 166-176.

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. Language, Cognition and Neuroscience, 1-13.

Broadbent, D.E.; Ladefoged, P. & Lawrence, W. (1956) Vowel Sounds and Perceptual Constancy, *Nature* **178**, 815–816.

Chandrasekaran, B., Chan, A. H., & Wong, P. C. (2011). Neural processing of what and who information in speech. Journal of cognitive neuroscience, 23(10), 2690-2700.

Chistovich, L.A., Sheikin, R.L. and Lublinskaja, V.V. (1979) “Centres of gravity” and spectral peaks as the determinants of vowel quality. In B. Lindblom and S. Öhman (Eds.) *Frontiers of Speech Communication Research*. New York: Academic Press (pp. 143-157).

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics*, 80(3), 784-797.

Creutzfeldt O, Ojemann G, Lettich E (1989a) Neuronal activity in the human lateral temporal lobe. I. Responses to speech. *Exp Brain Res* **77**, 451–475

Dabbs, J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men. *Personality and Individual Differences*, **27**(4), 801-804.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, 380(6574), 499.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience*, 23(8), 3423-3431.

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology,* 85(3), 1220-1234.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, 109(8), E505-E514.

Diesch, E., & Luce, T. (1997). Magnetic fields elicited by tones and vowel formants reveal tonotopy and nonlinear summation of cortical activation. *Psychophysiology*, 34(5), 501-510.

Diesch, E., & Luce, T. (2000). Topographic and temporal indices of vowel spectral envelope extraction in the human auditory cortex. *Journal of cognitive neuroscience*, 12(5), 878-893.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11), 1664-1670.

Disner, S. (1980) Evaluation of vowel normalization procedures, *J. Acoust. Soc. Am.* **67**, 253–261.

Drane, D. L., Ojemann, J. G., Phatak, V., Loring, D. W., Gross, R. E., Hebb, A. O., ... & Barsalou, L. (2013). Famous face identification in temporal lobe epilepsy: support for a multimodal integration model of semantic memory. *Cortex*, 49(6), 1648-1667.

Edmonds, B.A.; James, R.E.; Utev, A.; Vestergaard, M.D.; Patterson, R.D. & Krumbholz, K. (2010) Evidence for early specialized processing of speech formant information in anterior and posterior human auditory cortex. *European J. Neurosci*. **32**, 684-692. doi:10.1111/j.1460-9568.2010.07315.x

Evans, B. & Iverson, P. (2004) Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *J. Acoust. Soc. Am.* **115**, 352–361.

Fabricius, A., Watt, D. & Johnson, D.E.. (2009) A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change* **21**, 413-35.

Fant, G. (1975) Non-uniform vowel normalization. *STL-QPSR 2-3/1975*, 1-19.

Ferrand, C.T. (2002) Harmonics-to-noise ratio: An index of vocal aging. *J Voice* **16**, 480–487.

Fitch W.T. (1997) Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J. Acoust. Soc. Am.* **102**, 1213–1222. doi: 10.1121/1.421048 PMID: 9265764

Fitch, W. T. and Giedd, J. **(**1999**)** Morphology and development of the human vocal tract: A study using MRI. *J. Acoust. Soc. Am.* **106**, 1511-1522.

Fitzpatrick, D. C., Kuwada, S., Kim, D. O., Parham, K. & Batra, R. Responses of neurons to click-pairs as simulated echoes: auditory nerve to auditory cortex. J. Acoust. Soc. Am. 106, 3460–3472 (1999).

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, 40(4), 859-869.

Formisano, E.; De Martino, F.; Bonte, M.; & Goebel, R. (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* **322**, 970–973.

Fox, R. A., and Qi, Y.-Y. (1990). "Context effects in the perception of lexical tone," *J. Chin. Linguist.* **18**, 261–284.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, 119(3), 1712-1726.

Garau, G., Renals, S., Hain, T., 2005. Applying vocal tract length normalization to meeting recordings. In: *Proceedings of Interspeech* 2005, pp. 265–268. Lisbon.

Gerstman, L. (1968) Classification of self-normalized vowels, *IEEE Trans. Audio Electroacoust*. **AU-16**, 78–80.

Guilherme, J.M.; Garcia, E.W; Tewksbury, B.A.; Wong, B.A. & Kimbell, J.S. (2009) Interindividual variability in nasal filtration as a function of nasal cavity geometry. *J. Aerosol Med. & Pulmon. Drug Delivery* **22**, 139-155.

Glaser, R., York, A., & Dimitrakakis, C. (2016). Effect of testosterone therapy on the female voice. *Climacteric*, **19**(2), 198–203. http://doi.org/10.3109/13697137.2015.1136925

Greenspan, S.L.; Nusbaum, H.C. & Pisoni, D.B. (1988) Perceptual learning of synthetic speech produced by rule. *J. Exp. Psych.: Learn., Mem. & Cog.* **14**(3), 421-433. http://dx.doi.org/10.1037/0278-7393.14.3.421

Harnsberger, J.D., Brown, W. S. Jr., Shrivastav, R. & Rothman, H. (2010) Noise and Tremor in the Perception of Vocal Aging in Males, *Journal of Voice* 24(5), 523 - 530.

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci*. **8**, 393–402.

Hillenbrand, J.; Getty, L.A.; Clark, M.J. & Wheeler, K. (1995) Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* **97**(5), 3099-3111.

Hindle, D. (1978) 'Approaches to formant normalization in the study of natural speech. In *Linguistic Variation, Models and Methods*, edited by D. Sankoff (Academic, New York).

Hollien, H. (2001) *Forensic Voice Identification*. NY: Academic Press.

Hollien, H., Green, R., and Massey, K. (1994). "Longitudinal research on adolescent voice change in males," *J. Acoust. Soc. Am.* **96**, 2646–2653.

Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., and Chang, E.F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *J. Neurosci.* **36**, 2014-2026.

Johnson, K. (1989) Higher formant normalization results from auditory integration of F2 and F3. *Perception and Psychophysics*. **46**, 174-180.

Johnson, K. (1990a) The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am. 88*, 642-654.

Johnson, K. (1990b) Contrast and normalization in vowel perception. *Journal of Phonetics*. **18**: 229-254.

Johnson, K. (1997) Speech perception without speaker normalization: an exemplar model. In K. Johnson and J.W. Mullennix (eds.) *Talker Variability in Speech Processing*. San Diego: Academic Press (pp. 145-166).

Johnson, K., Ladefoged, P. & Lindau, M. (1993) Individual differences in vowel production. *J. Acoust. Soc. Am. 94*, 701-714.

Johnson, K., Strand, E.A. and D'Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics 27*, 359-384.

Jongman, A., & Moore, C. (2000). The role of language experience in speaker and rate normalization processes. In *Proceedings of the 6th International Conference on Spoken Language Processing* (Vol. I, pp. 62-65).

Joos, M.A. (1948) Acoustic Phonetics, *Language 24, Suppl. 2*, 1-136.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4), 736.

Kiefte, M. & Kluender, K. R. (2008) Absorption of reliable spectral characteristics in auditory perception, *J. Acoust. Soc. Am.* **123**(1), 366–376.

Kinnunen, T. & Li, HZ (2010) An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication* **52**(1), 12-40.

Ladefoged, P. (1989) A note on “Information conveyed by vowels.” *J. Acoust. Soc. Am.* **85**, 2223-2224.

Ladefoged, P., and Broadbent, D. E. (1957). “Information conveyed by vowels,” *J. Acoust. Soc. Am.* **39**, 98–104.

Lammert, A.C. & Narayanan, S.S. (2015) On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLoS ONE* **10**(7): e0132193. https://doi.org/10.1371/journal.pone.0132193

Latinus, M.; Crabbe, F.; & Belin, P. (2011) Learning-induced changes in the cerebral processing of voice identity. *Cereb. Cortex* **21**, 2820–2828.

Leather, J. (1983). “Speaker normalization in perception of lexical tone,” *J. Phonetics* **11**, 373–382

Lee, C. Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics,* 37(1), 1-15.

Liberman, M.Y. & Church, K. (1992) Text analysis and word pronunciation in text-to-speech synthesis. In Furui and Sondhi, Eds., *Advances in Speech Technology*, Marcel Dekker. (pp. 791-832)

Liebenthal, E., Desai, R., Humphries, C., Sabri, M., & Desai, A. (2014). The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. *Frontiers in neuroscience*. 8, 289.

Lloyd, R.J. (1890a) Some Researches into the Nature of Vowel-Sound. (Turner & Dunnett, Liverpool, England).

Lobanov, B. M. (1971) ‘Classification of Russian vowels spoken by different speakers, *J. Acoust. Soc. Am.* **49**, 606–608.

Maddux, S.D.; Butaric, L.N.; Yokley, T.R. & Franciscus, R.G. (2017) Ecogeographic variation across morphofunctional units of the human nose. *Am. J. Phys. Anthropol.* **162**, 103-119.

Magnuson, J.S. & Nusbaum, H.C. (2007) Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psych.: H. Percept. and Perf.* **33**(2), 391-409

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Mesgarani, N.; Cheung, C.; Johnson, K. & Chang, E.F. (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006-10.

Miller, J. D. (1989) Auditory-perceptual interpretation of the vowel, *J. Acoust. Soc. Am.* **85**, 2114–2134.

Miller, J.L. & Liberman, A.M. (1979) Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics* **25**(6), 457-465.

Miller, J. L., & Grosjean, F. (1981). How the components of speaking rate influence perception of phonetic segments. Journal of Experimental Psychology: Human Perception and Performance, 7(1), 208-215.

Miller, J. L., Aibel, L L., Green, K. (1984a). On the nature of rate-dependent processing during phonetic perception. Perception Psychophysics, 35, 5-15.

Miller, J. L., Green, K., & Schermer, T. M. (1984b). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. Perception & Psychophysics, 36(4), 329-337.

Miller, R.L. (1953) Auditory tests with synthetic vowels. *J. Acoust. Soc. Am. 25*, 114-121.

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* **32**, 14205-14216.

Monahan, P.J. & Idsardi, W. J. (2010) Auditory Sensitivity to Formant Ratios: Toward an Account of Vowel Normalization. *Lang Cogn Process* **25**: 808–839. doi: 10.1080/01690965.2010.490047

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *J. Acoust. Soc. Am* **102**, 1864-1877.

Morrill, T. H., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. *Cognition*, 131(1), 69-74.

Myers, E.B. & Theodore, R.M. (2017) Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language* **165**, 33-44.

Näätänen, R. & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. Psychological Bulletin, 125(6), 826–859.

Nagel, K.I. & Doupe A.J. (2008) "Organizing Principles of Spectro-Temporal Encoding in the Avian Primary Auditory Area Field L" *Neuron*, **58**(6):938-955

Nearey, T. M. (1978) *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club. Bloomington, Indiana.

Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. Journal of phonetics, 37(1), 46-65.

Nordström, P.E. and Lindblom, B. (1975): A normalization procedure for vowel formant data, paper 212 at the International Congress of Phonetic Sciences. Leeds, England.

Nott, G. (2018) The ATO now holds the voiceprints of one in seven Australians. Computerworld https://www.computerworld.com.au/article/633461/ato-now-holds-voiceprint-one-seven-australians/ accessed Feb 22, 2018.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. *Speech perception, production and linguistic structure*, 113-134.

Obleser, J., Elbert, T., Lahiri, A., Eulitz, C. (2003) Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Brain Research* **15**(3):207–213.

Obleser, J., Eisner, F. (2009) Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*. **13**(1):14–19.

Ohl, F.W. & Scheich, H. (1997) Orderly cortical representation of vowels based on formant interaction, *Proc. Natl. Acad. Sci*. USA **94**, 9440–9444.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience,* 18(6), 903-911.

Peng, G., Zhang, C., Zheng, H. Y., Minett, J. W., & Wang, W. S. Y. (2012). The Effect of Intertalker Variations on Acoustic–Perceptual Mapping in Cantonese and Mandarin Tone Systems. *Journal of Speech, Language, and Hearing Research*, 55(2), 579-595.

Pérez-González, D. & Malmierca, M. S. Adaptation in the auditory system: an overview. *Front. Integr. Neurosci.* 8, 1–10 (2014).

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., ... & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, 119, 164-174.

Peterson, G.E. (1951) The phonetic value of vowels. *Language* **27**, 541-553.

Peterson, G.E. (1961) Parameters of vowel quality. *Journal of Speech and Hearing Research 4*, 10-29.

Peterson, G.E. & Barney, H.L. (1952) Control methods used in the study of vowels. *J. Acoust. Soc. Am. 24*, 175-184.

Phillips, E. A. K., Schreiner, C. E. & Hasenstaub, A. R. Cortical Interneurons Differentially Regulate the Effects of Acoustic Context. Cell Rep. 20, 771–778 (2017).

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics,* 78(1), 334-345.

Potter, R. & Steinberg, J. (1950) Toward the specification of speech. *J. Acoust. Soc. Am. 22*, 807-820.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage, 62(2), 816-847.

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H. & King, A. J. Contrast Gain Control in Auditory Cortex. Neuron 70, 1178–1191 (2011).

Rauschecker, J.P. & Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* **12**, 718-724.

Rauschecker, J.P. & Tian, B. (2000) Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proc. Nat. Ac. Sci, USA*. **97**, 11800-6.

Reby, D. & McComb, K. (2003) Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behavior* **65**, 519-530

Reinisch, E., Jesse, A., & McQueen, J. M. (2011a). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. Language and Speech, 54(2), 147-165.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011b). Speaking rate from proximal and distal contexts is used during word segmentation. Journal of Experimental Psychology: Human Perception and Performance, 37(3), 978.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. Journal of Phonetics, 41(2), 101-116.

Roberts, T. P. L., Flagg, E. J. & Gage, N. M. (2004). Vowel categorization induces departure of M100 latency from acoustic prediction. Neuroreport, 15(10), 1679–1682.

Saenz, M., and Langers, D.R.M. (2014). Tonotopic mapping of human auditory cortex. *Hear. Res*. **307**, 42-52.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS computational biology*, 10(1), e1003412.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., & Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences*, 114(18), 4799-4804.

Sawusch, J. & Newman, R. (2000)

Schwippert, C. and Benoit, C. (1997) Audiovisual intelligibility of an androgynous speaker. In *Proceedings of the ESCA workshop on audiovisual speech processing (AVSP'97): Cognitive and computational approaches*, Rhodes, Greece (C. Benoit & R. Campbell, editors), pp. 81-84.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. Trends in neurosciences, 26(2), 100-107.

Shestakova, A., Brattico, E., Soloviev, A., Klucharev, V., & Huotilainen, M. (2004). Orderly cortical representation of vowel categories presented by multiple exemplars. *Cognitive Brain Research*, 21(3), 342-350.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011a). Constraints on the processes responsible for the extrinsic normalization of vowels. Attention, Perception, & Psychophysics, 73(4), 1195-1215.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. Neuropsychologia, 49(14), 3831-3846.

Sjerps, M., McQueen, J. M., & Mitterer, H. (2012). Extrinsic normalization for vocal tracts depends on the signal, not on attention. In Thirteenth Annual Conference of the International Speech Communication Association.

Sjerps, M. J., McQueen, J. M., & Mitterer, H. (2013). Evidence for precategorical extrinsic vowel normalization. Attention, Perception, & Psychophysics, 75(3), 576-587.

Sjerps, M. J., & Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. Journal of Phonetics, 41(3-4), 145-155.

Slawson, A.W. (1968) Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *J. Acoust. Soc. Am.* **43**, 87-101.

Smith, D.R.R. & Patterson, R.D. (2005) The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *J. Acoust. Soc. Am.* **118**(5), 3177; doi:10.1121/1.2047107

Stilp, C.E.; Alexander, J.M.; Kiefte, M.J. & Kluender, K.R. (2010) Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception & Psychophysics* **72**, 470–480.

Subramaniam, R.P; Richardson, R.B; Morgan, K.T.; Kimbell, J.S. & Guilmette, R.A. (1998) Computational fluid dynamic simulations of inspiratory airflow in the human nose and nasopharynx. *Inhalation Toxicology* **10**, 91-120.

Suga, N., O'Neill, W.E., Kujirai, K. and Manabe, T. (1983) Specificity of combination-sensitive neurons for processing of complex biosonar signals in auditory-cortex of the mustached bat, *Journal of Neurophysiology*, **49,** 1573–1627)

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1074.

Summerfield, Q.; Haggard, M.; Foster, J. & Gray,S. (1984).Perceiving vowels from uniform spectra: Phonetic exploration of an auditory after effect. *Perception & Psychophysics* **35**(3), 203–213.

Sussman, H.M. (1986) A neuronal model of vowel normalization and representation. *Brain Lang. **28**,* 12-23.

Syrdal, A. K., and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels, *J. Acoust. Soc. Am*. **79**, 1086–1100.

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science*, 357(6353), 797-801.

Thomas, E.R. (2011) *Sociophonetics: An Introduction*. New York: Palgrave MacMillan

Toscano, J.C., McMurray, B., Dennhardt, J. & Luck, S.J. (2010). Continuous perception and graded categorization electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. Psychological Science, 21(10), 1532–1540.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, cognition and neuroscience*, 30(5), 529-543.

Traunmüller, H. (1981) Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am. 69*, 1465-1475.

Traunmüller, H. (1984) Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Communication 3,* 49-61.

Turkeltaub, P. E., & Coslett, B. H. (2010). Localization of sublexical speech perception components. Brain and language, 114(1), 1-15.

Ulanovsky, N., Las, L., Farkas, D. & Nelken, I. Multiple Time Scales of Adaptation in Auditory Cortex Neurons. J. Neurosci. 24, 10440–10453 (2004).

Verbrugge, R.R., Strange, W., Shankweiler, D.P. & Edman, T.R. (1976) What information enables a listener to map a talker's vowel space? *J. Acoust. Soc. Am. 60*, 198-212.

Von Kriegstein, K; Giraud, A.L. (2004) Distinct functional substrates along right superior temporal sulcus for the processing of voices. *NeuroImage* **22**, 948-955.

Von Kriegstein, K; Warren, J.D.; Ives, D.T.; Patterson, R.D. & Griffiths, T.D. (2006) Processing the acoustic effect of size in speech sounds. *NeuroImage* **32**, 368-375.

Wakita H. (1977) Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans Acoust Speech Sig Proc.* **25**, 183–192. doi: 10.1109/TASSP.1977.1162929

Waldron, E. J., Manzel, K., & Tranel, D. (2014). The left temporal pole is a heteromodal hub for retrieving proper names. Frontiers in bioscience (Scholar edition), 6, 50.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, 57(8), 1124-1133.

Watkins, A.J. (1988) Spectral transitions and perceptual compensation for effects of transmission channels. In *Proceedings of the 7th Symposium of the Federation of Acoustical Societies of Europe: Speech '88*. W. Ainsworth & J. Holmes (Eds). Edinburgh, UK: Institute of Acoustics.

Watkins, A.J. (1991) Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* **90**, 2942-2955.

Watkins, A.J. & Makin, S.J. (1994) Perceptual compensation for speaker differences and for spectral-envelope distortion. *J. Acoust. Soc. Am.* **96**, 1263-1282.

Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter-and intratalker variation in Cantonese level tones*. Journal of Speech, Language, and Hearing Research*, 46(2), 413-421.

Woolley, S. M., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature neuroscience*, **8**(10), 1371.

Young, E. D. (2008). Neural representation of spectral and temporal information in speech. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363(1493), 923-945.

Zäske, R., Hasan, B. A. S., & Belin, P. (2017). It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex*, 94, 100-112

Zhang, C., Peng, G., & Wang, W. S. (2012). Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. *The Journal of the Acoustical Society of America*, 132(2), 1088-1099.

Zhang, C., Peng, G., & Wang, W. S. Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and language*, 126(2), 193-202.

Zhang, C.C. & Chen, S. (2016) Toward an integrative model of talker normalization. *J. of Exp. Psych.: Hum. Perc. and Perf.*, **42** (8), 1252-1268.

Zimman, Lal (2017). Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/. *Language in Society* **46**(3):339-370.