Not Published
Vol. –

ISSN XXX-XXXX-XXXX

Speaker and Dialect Normalization in Speech Perception: ERP Evidence

Cristina Cumpănăşoiu

Leiden University Centre for Linguistics

Master Thesis Coordinator: Prof. Dr. Niels O. Schiller (Leiden University)

Master Thesis Co-reader: Dr. Kateřina Chládková (University of Amsterdam)

Abstract

Speaker normalization is a process during speech perception through which the vocal tract variabilities of different speakers are minimized while preserving the phonemic and sociolinguistic variation, prior to the recognition of linguistic categories. This study aims at deciphering the underlying mechanisms through which listeners are able to cope with speaker and dialect differences. Using an event-related potential (ERP) oddball experiment, the present study examined whether listeners treat between-speaker variability in vowel acoustics differently than they treat between-dialect variability. In contrast to the results of a previous experiment, results from the present ERP study show a higher mismatch negativity (MMN) for gender variation than for speaker changes indicating that listeners do not normalize gender differences while changes in speaker are more easily normalized.

Keywords: speaker normalization, speech perception, ERP

Introduction

Speech is an acoustic signal that originates from the human vocal tract which alongside with the message transmitted also performs a variety of additional functions such as indicating physical and sociolinguistic characteristics of the speaker. The characteristic speech signal is divided into two halves: the first one encrypts speech information and the second half consists of silent or noise parts and it is represented by the data between the verbal utterances (Cutler & Blumstein, 2003). The activity of speaking involves generating a noise with the body, shaping and articulating the noise into meaningful sounds. Sounds are produced by pressing air through the glottis and the controlled tension of the vocal folds produces the opening and closing of the cords around the column of air being pushed up from the lungs. The vocal folds vibrate to generate a sound and they also modulate the volume and pitch of the sound. For the production of unvoiced speech, the air exhaled out of the lungs and through the trachea in not affected by the vibrating vocal folds (Collins & Mees, 2003).

The verbal or voiced speech is made up mostly of vowel sounds (O'Connor, 2015). Vowels are sounds produced when the vocal folds vibrate as a reaction to a movement of air. Different positions of the lips and tongue imply changes to the oral cavity which in turn will result in a different resonance. Therefore, vowel qualities are produced. When vowel sounds are produced formants are adjusted within a set interval of frequencies. The formant frequencies are determined by the shape of the vocal tract (Sundberg, 1977).

The final result, the speech signal, reflects information regarding the message, the speaker, the language but also information regarding the emotional condition of the speaker. While this variability does not modify the semantic content of the utterances, it can definitely provide information about the speakers' input for the complete perception of the message (Krishna, Patil, & Elhilali, 2012).

Since the invention of the spectrograph, questions regarding the perception of speech started to arise as variability within and between speakers was observed. Speech perception studies are concerned with the process by which linguistic information is extracted by

listeners from a highly variable acoustic input. The process is even more complex considering that spoken language is variable in its production and highly stable during its perception (Krauss & Pardo, 2006). It is exactly this high variability in the production of speech but also the stability of its perception that have attracted the interest of psycholinguists over the last decades. Within and between speaker variation in acoustic cues has been closely examined since more than five decades. Speaker normalization in speech perception focuses on the acoustic variation between speakers when utterances are phonologically identical, and examines the ability that listeners have to identify words uttered by different speakers despite this variability. The human brain receives through the hearing mechanism a neural pattern which is related to the spectral envelope of the vowel spectrum. This curve in the frequency-amplitude plane differs depending on the pitch and the timbre of the sound (Bolt et al., 1970). The process through which listeners transform the quality of a vowel into the correspondent standard vowel quality is called normalization.

The spectrogram in figure 1 (Wood, 2015) shows the same vowel produced by different speakers. The values of the formants for the same vowel uttered by different speakers vary. While the first formant for the word uttered by the first speaker (male) indicates a lower formant values in comparison to the second one (female) which shows higher formant values. Despite these variabilities, listeners are still able to perceive the vowel as being the same one without prior exposure to any of the speakers (Hallé & Boysson-Bardies, 1994). Research on speech normalization attempts to explain how listeners are able to extract speech from a permanently variable signal coming from different sources.

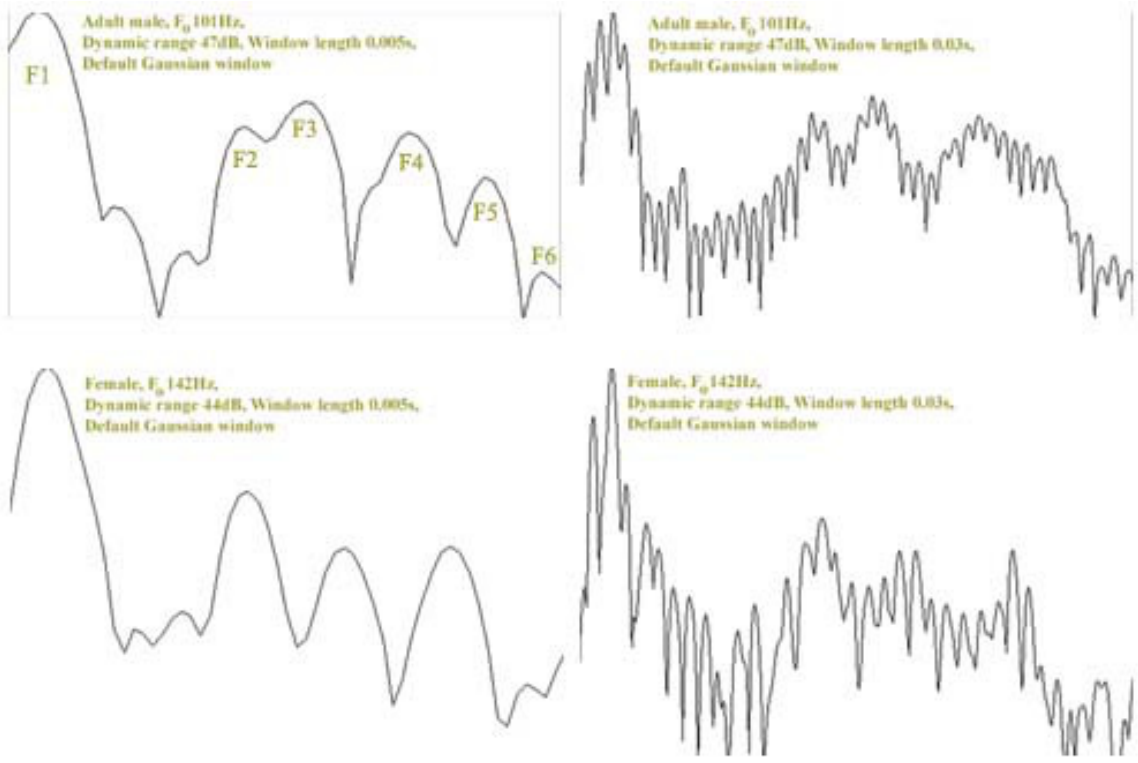

Figure 1. Male and female formants for the vowel [i] (Wood, 2015)

The Perception of Vowels

Vowels represent an important characteristic in spoken languages as all languages distinguish themselves through the number of vowels and through the unique acoustic properties of their vowels (Rosner & Pickering, 2008). Vowel formant frequencies vary upon the position of the tongue. Usually, the frequency of first formant (F1) has an inverse value to the height of the tongue. Thus, F1 frequency for /i/ and /u/ is low and they are produced with a high tongue position. The F1 frequency for /a/ is high and the vowel is produced with a low tongue position. The absolute formant values of every vowel are different for every speaker (Lehiste & Meltzer, 1973). The first formant(F1) values for /i/ may vary from speaker to speaker from 180Hz to 400Hz and the second formant (F2) between 2000Hz and 3500Hz.When talking about vowel perception, identification and constancy make up the two most important issues. The input of vowel formants in vowel perception has been demonstrated over the last decades by numerous researchers. Early studies using synthetic speech

(Ladefoged & Broadbent, 1957) show that the process of vowel identification contains a normalization stage during which listeners size their perceptual apparatus for each speaker's vowel space. The same studies indicate that raising or lowering the formants during an introductory sentence affects the identification of the vowel belonging to the following test word. Fry, Abramson, Eismas, and Liberman (1962) used a continuum of synthetic vowels for a study which demonstrates that the main factors during vowel recognition are the positioning of the first two formants. The perceptual influence of higher formants were also demonstrated. Studies done by Fujisaki and Kawashima (1968) indicate the effect of F3 with two different vowel continua. A vowel category boundary shift of 200 Hz between F1 and F2 for a /u/-/e/ continuum was produced by an F3 shift of 1500 Hz and for an /o/-/a/ continuum a boundary shift of only 50 Hz was obtained.

The role of the fundamental frequency(F0) in the perception of vowels has also been demonstrated. Slawson (1968) reported that F1 and F2 frequencies are increased by a small rate if F0 is increased from a standard male value to a standard female value demonstrating the influence of the fundamental frequency in vowel quality. The effect of F0 on the values of the first two formants within-speakers was also investigated and the results indicate that when speakers speak at a high F0, they raise their formants at the same time. The effect for the F1 changes is more noticeable on women than on men (Chládková, Boersma, & Podlipský, 2009). Similar research on the influence of F0 in perception indicates that for F0 shifts of 200 Hz boundary shifts by 100 Hz up to 200 Hz for F1 (Fujisaki & Kawashima, 1968). Gottfried and Chew (1986) could demonstrate that the process of identification of vowels was hardened when vowels were produced by a counter tenor at a much higher F0 than what it is normal for a male voice. Investigations on the perception of synthetic vowels with F0 up to 700 Hz specific mainly for children were performed by Traunmüller (1981) concluded the influence of the fundamental frequency on vowel perception.

The role of duration in vowel perception has also been examined. Studies on English vowels involving synthesized two-formant vowels with static formant frequencies varying in duration from 120 to 600 ms indicate that listeners' perception is affected by the duration

of vowels (Ainsworth, 1972). Ainsworth noticed that listeners have a high probability on identifying a vowel in the /u/-/ʊ/ spectrum as /u/ if the vowel is long or /ʊ/is the vowel is short. Along with that, results also show that high vowels such as /i/-/ɪ/ or /u/-/ʊ/ are less influenced by duration than analogous vowels with different duration. The findings were confirmed through similar results by Bennett (1968). Thereby, the perceptual influence of duration for vowels differs along vowel categories and the role of duration is conditioned by the spectral characteristics of a specific vowel.

Previous Experiments

Experiments investigating the underlying mechanisms behind speaker and dialect normalization, specifically studies which examined whether listeners handle between-speaker variability in vowel acoustics the same way they handle between dialect variability were performed on Australian English monolinguals and bilinguals (Dadwani, Peter, Chladkova, Geambasu, & Escudero, 2015) and on Dutch speakers (Chládková et al., in preparation). The studies consisted of an event-related potential (ERP) experiment which allowed for a thorough analysis of the online processing of speaker and dialect variation in vowel acoustics. The experiment investigated the way listeners attend to 4 different types of changes: vowel, speaker, gender and dialect changes. The hypotheses stated that if listeners normalize speaker and dialect changes alike, they would neglect variability in voices and dialects. If listeners do not normalize speaker and dialect alike, they would normalize variability in isolated vowels between–speakers but not between–dialects. The stimuli used for both experiments were natural tokens of Dutch vowels /ɪ/ and /ɛ/ extracted from monosyllabic words. The five different stimuli used were one female speaker's ND /ɪ/ and /ɛ/, a different female speaker's ND /ɪ/, a male ND /ɪ/, and a female SD /ɪ/. The experiment consisted of a multi-deviant oddball paradigm in which a frequently repeated standard was interspersed by infrequent repetitions of four different deviant stimuli (Dadwani et al., 2015).

From the ERP data the amplitude of the mismatch negativity (MMN) and the amplitude of the P3a component were studied. Results from the Dadwani et al. (2015) study

indicated that Australian English listeners show sensitivity to accent changes and are less sensitive to vowel changes than they are to speaker variation. Results from the study performed on Dutch speakers showed similar results. Listeners do not automatically normalize dialect differences and they normalize more readily changes in speaker than changes in gender (Chládková et al., in preparation). As the similarity between the results from both studies was believed to be a result of the strong difference in voice qualities between the male and female speakers a follow up study was performed in order to confirm if the acoustic differences in voice quality (namely F0) between the deviants and the standard influenced the results for the both experiments.

The Follow-up Study: Motivations and Outline

Motivations. The stimuli used for the present study were chosen based on the premise that voice quality is a salient cue in an unattended discrimination task (Dadwani et al., 2015), hence the difference between the stimuli used in the preceding experiments and the present one. Due to the large MMNs triggered when stimuli differed in F0 from the standard (see table 1) indicating that listeners could have treated deviant stimuli as variants of the standard, the stimuli used in the present study were vowels manually corrected to be approximately 60 ms per stimulus and with a voice quality between deviants and the standard comparable for all deviants. Using the new set of stimuli, could lead to different results in comparison to the preceding studies. A different outcome would indicate that the similarity between the results from the preceding studies was due to the differences in fundamental frequency.

deviant type	difference from standard			
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)
standard	177	395	2306	2773
dialect	212	520	1854	2942
gender	136	317	1775	2325
speaker	176	424	2289	2982
vowel	178	573	1960	2862

Table 1
Difference between stimuli and standard used in Chládková et al. (in preparation)

Outline. This experiment is a follow-up of the studies done by Chládková et al. (in preparation) and Dadwani et al. (2015) which explore the mechanisms behind speaker versus dialect normalization. Fig. 1 illustrates this variability with Dutch vowels (Adank, Van Hout, & Van de Velde, 2007) and shows the average first and second formant characteristics of /ɪ/, /ɛ/, /u/, and /ɔ/ produced by female and male speakers from North Holland and East Flanders. Two main aspects can easily be observed in the figure. The first one is represented by the considerable differences between the vowels produced by a man and those produced by a woman. The second important aspect to be noticed is the difference for the front vowels /ɪ/ and /ɛ/ between the two dialects.

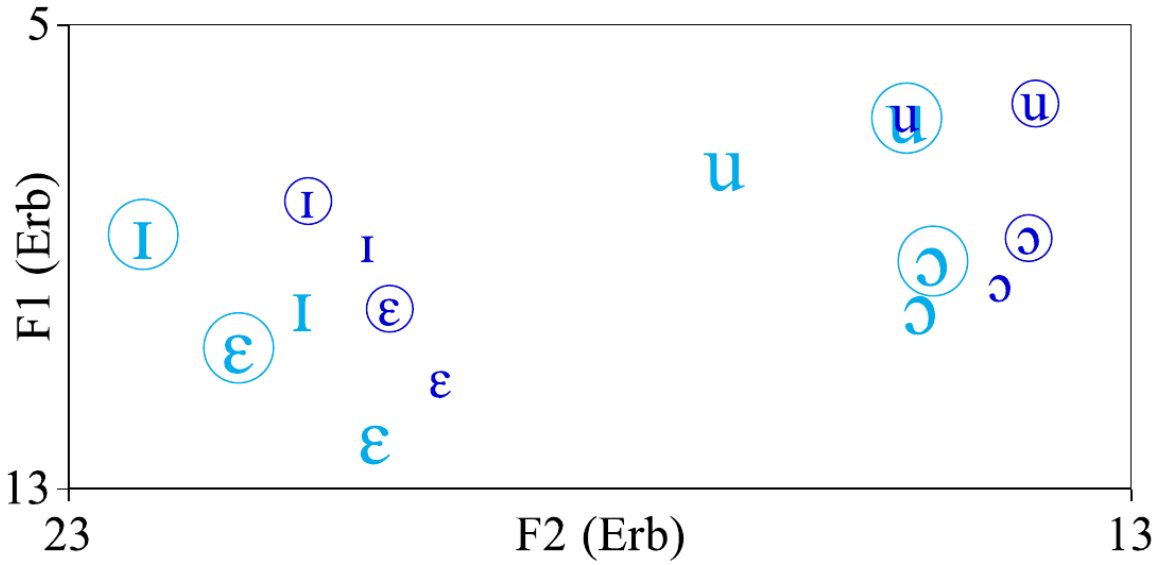

Figure 2. Dutch vowels /ɪ/, /ɛ/, /u/, /ɔ/ : average F1 and F2. Large light symbols: vowels produced by women. Small dark symbols: vowels produced by men. Circled symbols: vowels from North Holland. Plain symbols: vowels from East Flanders (Dadwani et al., 2015).

Based on these premises, the main focus of the present study is to investigate whether listeners cope with between-speaker variation in a similar way they cope with between–dialect variation. For this study isolated vowels rather than words were used to avoid the activation of lexical knowledge, since dialect and speaker normalization may appear to be similar in the context of familiar words.

The initial hypotheses of the experiment were that listeners should ignore variability

in voices and dialects when classifying vowel tokens if they normalize speaker and dialect similarly. The second hypothesis was that if speaker and dialect normalization do not undergo the same process, listeners should have the ability to normalize variability in isolated vowels between–speakers but not between–dialects. Due to the modifications of the stimuli regarding the voice quality between the deviants and the standard used in the present study, the hypotheses can now be more appropriately addressed. The preceding studies consisted of an event-related potential (ERP) experiment which allowed for a thorough analysis of the online processing of speaker and dialect variation in vowel acoustics.

The prediction was that listeners would adapt to speaker variability in isolated vowels classifying different dialects as different vowel categories, and that different dialects would result in a mismatch response similar to the response of the different vowel categories. Recall that results from the previous studies showed that listeners do not automatically normalize dialect differences. Moreover, changes in gender showed a smaller MMN response than changes in speaker indicating the fact that listeners normalize changes in gender more easily than differences in speaker.

The methodological approach: Mismatch Negativity (MMN) and P3 Brain Potentials

A remarkable tool in understanding the mechanisms behind cognitive processes, event-related potentials (ERPs) represent a noninvasive technique by which the electrical activity of the brain as a result of a specific stimulus (sensory, cognitive or motor event). ERPs are measured using electroencephalography (EEG), a technique which records the electrical activity of the human brain by placing a determined number of electrodes on the scalp, amplifying the signal and plotting the changes in voltage over time (Berger, 1929) by means of an averaging method (Luck, 2014). By means of an averaging process brain activity unrelated to the stimuli is filtered out. The specific brain responses related to the stimuli are known as event-related potentials as a way to indicate that are electrical potentials corresponding to specific events. ERP waveforms represent a number of positive and

negative voltage deflections each of them with a different polarity, amplitude and duration which reflect a specific neural or psychological process (Kappenman & Luck, 2011).

The mismatch negativity (MMN) component of the event-related potential (ERP) is a brain reaction to violation of a rule as a consequence of a sequence of stimuli particularly in the auditory domain (Näätanen, 1992). While the MMN has been widely used as a means to study preattentive processing and storage of regularities in basic physical stimulus features, more recent studies involving auditory analysis reflected by MMN reveal as well the use of complex regularities such as the connection between different physical features of the stimuli or seven in patterns found in the auditory stream. The violation of these regularities elicits the MMN (Paavilainen, 2013). When electrophysiological techniques such as electroencephalography (EEG) or magneto encephalography (MEG) are employed, the MMN is obtained after subtracting the event-related response to the standard event from the response to the deviant event (Garrido, Kilner, Stephan, & Friston, 2009). The MMN has been widely used in neurolinguistics particularly in studies focusing on phonological or syntactic processing.

The P300 wave represents a positive centro-parietal deflection elicited during the process of sensory discrimination of a participant. When recorded by electroencephalography the P300 component peaks between 200 and 250ms or at a later stage (e.g. 400 up to 800 ms) depending on the difficulty of discrimination (Picton, 1992). Apart from its latency dependent on stimulus evaluation timing, the P3 component also varies according to the cognitive abilities of each participant (Polich, 2007). However, the P300 is not an unitary ERP component; two subcomponents i.e. P3a and P3b are identified. While the P3a is elicited at fronto-central electrodes (e.g. FCz, Cz) with a peak latency of 250 to 280 ms, the P3b emerges from temporal-parietal activity, has a peak latency between 250 and 500 ms and is associated with attention and also to succeeding memory processing. Experimental paradigms such as selective attention tasks or specific memory assignments in which participants are required to pay attention and evaluate stimuli, will elicit a P3b subcomponent. The amplitude varies according to the percentage of targets relative to the number

of standards as well as to the type of presentation and the frequency of the stimuli.

Methods

Participants

Four native speakers of Dutch, two males and two females took part in the study; age-range 20-37 years. All of the participants were recruited from Leiden University and participated in the experiment in exchange for book vouchers. All of them were right handed and reported normal hearing and no language or neurological impairments. Before testing, practical information regarding the experiment they would be subjected to was given to each participant while theoretical aims of the study were not revealed. They were instructed to read the instructions carefully and ask about any doubts they might have before starting testing. A consent form by which they agreed with all terms and conditions was signed by all participants prior to the experiment. Testing took place at Leiden University in The Netherlands. All participants were tested individually during a single session in a sound-proof room. Each of them was seated in a chair at a distance of one meter from the screen and were instructed not to blink or to move excessively in order not to introduce noise into the EEG data. While participants were watching a silenced movie with subtitles in Dutch, stimuli with a loudness of 65 dB were presented through two loud speakers placed at equidistant distances. Participants were told to ignore the sounds they would hear from the loud speakers.

The experiment was divided in three parts each followed by a short break in which the experimenter checked if everything was going alright with the participant. Apart from the four participants, another participant took part in the experiment, however, due to technical failure during recording the data could not be used.

Stimuli and Oddball Paradigm

Figure 4 shows the stimuli used in the present ERP study. All vowels used in the experiment were isolated naturally produced tokens of Dutch vowels /ɪ/ and /ɛ/ from the

corpus of Adank et al. (2007). In the figure the circled vowels are identified as the standard stimuli while the ones which are not circled represent the deviant stimuli. The intended vowels are represented by the IPA symbols while subscripts (speaker, gender, dialect, vowel) point out the type of change between the stimuli. Formant ratios are also indicated in the figure. They indicate the vowel positions the way they would look like if people normalized through ratios. Sensitivity to formant ratios could show the automaticity of speaker and gender normalization (Kriengwatana, Escudero, & Terry, 2014).

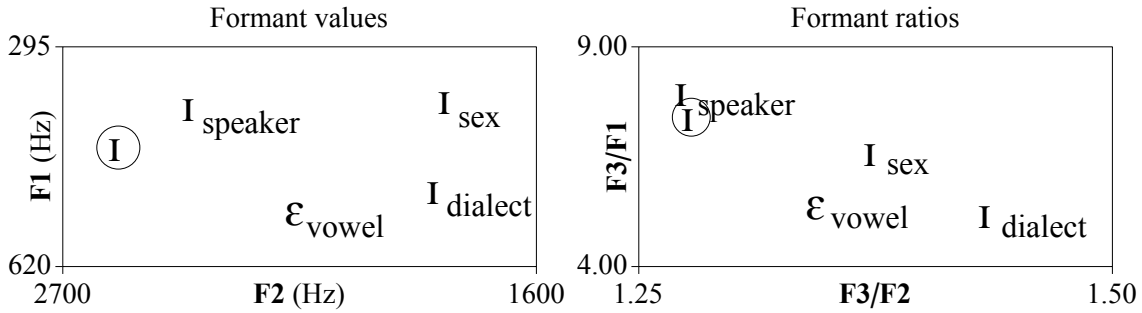

Figure 3. Standard and deviant stimuli

The new set of stimuli used in the present study were extracted from the Adank et al. (2007) corpus. The stimuli were specifically selected so that the voice quality between each deviant and the standard are comparable for all deviant types. The vowels were extracted from monosyllabic words /sɪs/ and /sɛs/. Only the central stable portion of the vowel was extracted so that any formant transitions of the flaking consonants were removed. After the tokens were selected they were judged by a Dutch-speaking phonetician as representative of the intended vowel category and dialect. Five different stimuli were selected: one female speaker's Northern Dutch (ND) /ɪ/ and /ɛ/, a different female speaker's ND /ɪ/, a male ND /ɪ/, and a female Standard Dutch (SD) /ɪ/. For the present study, the duration of the extracted vowels was manually corrected to be approximately 60 ms per stimulus, by either removing additional periods from the vowel's edges or duplicating some of the central periods. The intensity of the stimuli was equalized and ramped at the vowel edges (5-ms onset and offset portions).

If subjects normalize automatically vocal tract differences, speaker and gender deviant stimuli would produce a smaller MMN as a reaction to the fact that they are the closest to the standard. Based on the same premise, dialect and vowel deviant stimuli would yield a larger MMN due to the fact that they are further from the standard stimuli. Results from the preceding experiments showed that larger MMNs were triggered when stimuli differed in F0 from the standard (i.e. voice quality), which could indicate the fact the participants might have treated half of the deviants not as deviants but rather as variants of the standard. For the present study stimuli were selected so that the voice quality between each deviant and a standard are comparable for all deviant types. Table 2 lists the vowels' first three formants pitch and duration. Based on the MMN patterns elicited from the previous study and considering that for the present experiment deviant stimuli differ in F0 from the standard to the same extent, listeners' perception should not be affected by F0 differences between the stimuli. Table 3 shows the differences between the deviant and the standard stimuli.

stimulus	duration	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)
standard	59	211.02	444.18	2571.92	3285.55
dialect	59	243.07	507.31	1832.78	2628.23
speaker	60	249.80	384.27	2400.78	3058.56
gender	59	239.05	375.07	1805.86	2480.79
vowel	60	186.72	538.04	2158.74	2901.35

Table 2
Duration and first three formants of each of the stimuli

Each deviant type occurred in the oddball block and the ratio of presentation was 0.80 for the standard and 0.05 for each of the deviants, just like in the preceding experiments. The oddball block contained the same number of stimuli (i.e.2751) as in the previous experiments and had a total duration of approximately 35 minutes. Control blocks for all deviant types in which each deviant was presented 120 times, were introduced after the oddball block.

deviant type	difference from standard			
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)
dialect	-32.06	-63.13	739.14	657.32
gender	-38.78	59.91	171.14	226.99
speaker	-28.04	69.11	766.06	804.76
vowel	24.3	-93.86	413.18	384.2

Table 3
Difference between stimuli and standard used in the present study

EEG Recording and pre-processing

EEG was recorded from 64 active Ag-AgCl electrodes placed in a cap that was fitted to participant's head size (International 10/20 placement). For this experiment seven external electrodes were used. They were placed on the nose, below and above the right eye, on the left and right canthi, on the right and left mastoid. The EEG signal was recorded at 512 Hz. After all electrodes were placed correctly on the cap, the EEG acquisition software was opened and the experiment file was loaded. The activity of all electrodes was observed and for those which produced a flat line signal or showed very little or increased activity while the experiment had not started yet, more electrode gel was applied on the targeted channels. After ensuring that all electrode signals are functioning properly, testing and recording began. The EEG data was processed and analyzed in Praat 5.4(Boersma & Weenink, 2015), a software package for the analysis of speech and EEG signals in phonetics.

In order to avoid differences that could cause a change in the final results in the follow-up experiment parameters were kept the same as in the preceding the experiment (Chládková et al., in preparation) (Dadwani et al., 2015). In the present analysis the EEG signal was offline referenced to the mastoids. Before epoching and artifact removal, the EEG data was band-pass filtered in order to remove linear trends. As filters can distort the EEG data considerably forming artificial peaks and oscillations a low cut-off of 1 Hz was chosen and high cut-off of 30 Hz. Filters were applied to the continuous EEG and not to the epoched or averaged ERPs as filters work best with a continuous data (Luck, 2014). After filtering the data, in order to compare event-related EEG dynamics for two conditions of the same experiment, data was epoched from -100ms to +600ms relative to stimulus onset and

it was baseline corrected with respect to 100ms pre-stimulus interval for each participant. Before averaging, any data with amplitude surpassing +/- 75 µV were rejected. In the present experiment data of none of the participants exceeded 50% of artifact contamination which means that none of the data participants were rejected. Responses to each deviant and control stimulus type were averaged for each participant. For each participant, four difference waves were obtained by subtracting the responses to each control stimulus from their equivalent deviant. The resulting waves were grand-averaged across all participants. In the grand average difference waves, the latency of the negative peak was established within the time window 150 and 250 ms post stimulus onset (the MMN component) and the positive peak was determined between 200 and 400 ms post stimulus onset (the P3 component). Individually, a 40 ms window for each of the two grand peaks was set and the mean amplitude was measured in both 40 ms windows for every participant. The average voltage within the 40 ms windows were used as a measure for the MMN amplitude and P3 difference amplitude respectively.

Reproducible research

For the present study a set of scripts has been used in order to automatize the analysis and graphing of the experiment dataset. In order to reproduce all the intermediate data and charts from the original experiment data a number of scripts has been created. These scripts use praat (Boersma & Weenink, 2015), pandas (Lambda Foundry & Team, 2015)and matplotlib (Hunter & Team, 2015) in order to analyze, post-process and generate charts in a fully unattended manner. The scripts will also compute and report the reliability of the captured data. The code is available at hsversiottp://github.com/whirm/praat_eeg_scripts.

Results

This paragraph summarizes the results obtained from the previous studies: (Dadwani et al., 2015) and (Chládková et al., in preparation). The results showed that Australian English listeners are more sensitive to accent and gender changes and that they are less

sensitive to vowel variation than they are to speaker variation (Dadwani et al., 2015). The mean amplitudes in table 4 show the large mismatch response to accent in comparison to the much lower response to speaker changes, at channel FCz. Similar results were retrieved from the study performed on Dutch participants. Listeners do not automatically normalize dialect changes and they are more sensitive to gender variation than they are to speaker variation (Chládková et al., in preparation), see figure 4. The results submitted to two repeated measure ANOVAs showed a reliable MMN only for dialect and gender deviants, dialect eliciting a larger MMN than vowel deviant (p=.029) and the speaker deviant(p=.057). Gender yielded a larger MMN than speaker deviant (p=.019). P3 differences were also larger for dialect deviants than for the other three deviant types, see table 5.

Results from the present study will not be submitted to an ANOVA test. The data will be examined only visually as only four participants took part in the experiment.

deviant type	group	MMN amplitude
accent	monolinguals	-4.36 (0.93)
	bilinguals	-3.02 (0.43)
gender	monolinguals	-4.41 (1.15)
	bilinguals	-3.63 (0.63)
speaker	monolinguals	-2.68 (0.78)
	bilinguals	-2.36 (0.53)
vowel	monolinguals	-2.56 (0.72)
	bilinguals	-1.93 (0.41)

Table 4
MMN amplitudes for the four deviant types at channel FCz (Dadwani et al., 2015)

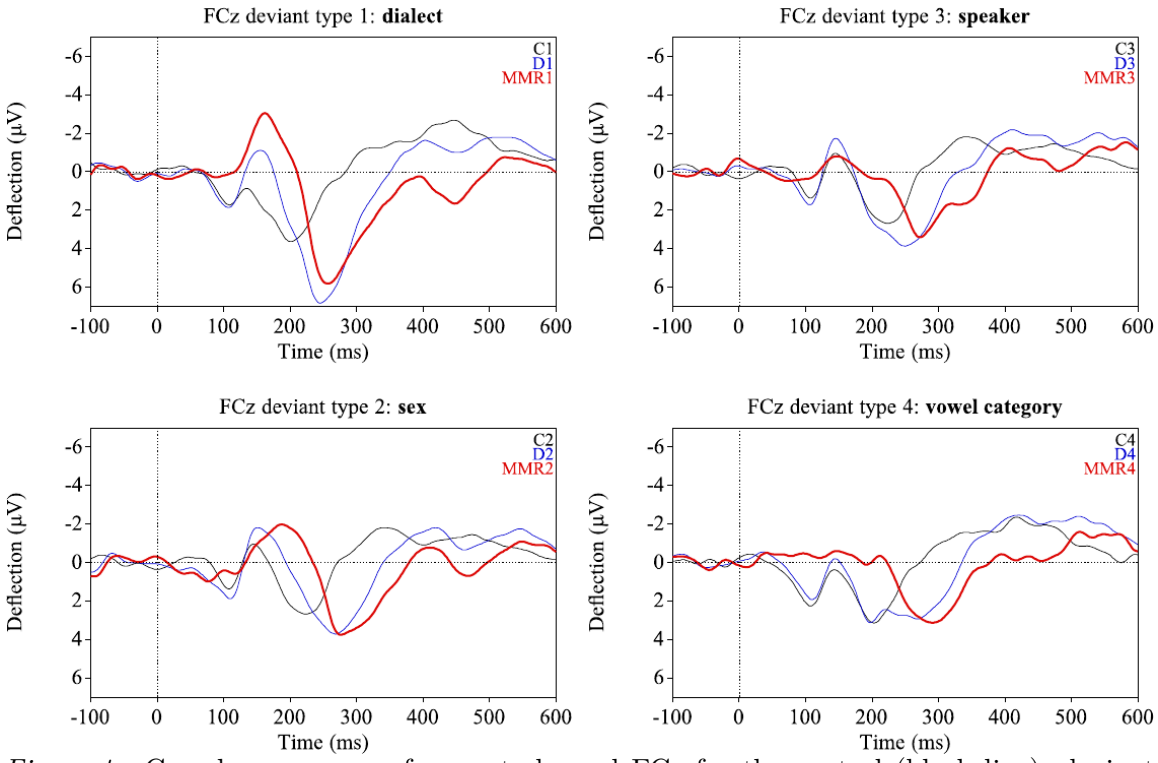

Figure 4. Grand-average waveforms at channel FCz for the control (black line), deviant stimuli (blue line) and difference waves (red line) per deviant type(Chládková et al., in preparation)

deviant type	mean MMN amplitude (95% c.i.)		mean P3a difference amplitude (95% c.i.)	
dialect	-2.390	(-3.816..-0.963)	4.485	(3.334..5.635)
gender	-1.796	(-2.973..-0.620)	2.672	(1.262..4.083)
speaker	-0.465	(-1.807..0.877)	2.540	(1.018..4.061)
vowel	-0.495	(-1.422..0.432)	2.596	(1.745..3.447)

Table 5
MMN amplitude and P3 differences (Chládková et al., in preparation)

Figure 5 plots the grand average waveforms of the deviant and control stimuli, as well as the average difference waves (i.e deviant – control) for each deviant type at channel Fz. The mismatch negativity event-related (MMN) potential was evaluated for all four participants who took part in the experiment. Channel Fz was chosen as a representative channel for analyzing the MMN response as the auditory-evoked MMN is usually strongest at frontal channels along the midline. The results were compared to the MMN responses elicited in the preceding experiments. For a more accurate comparison between the two

studies, results elicited from channels along the midline were used.

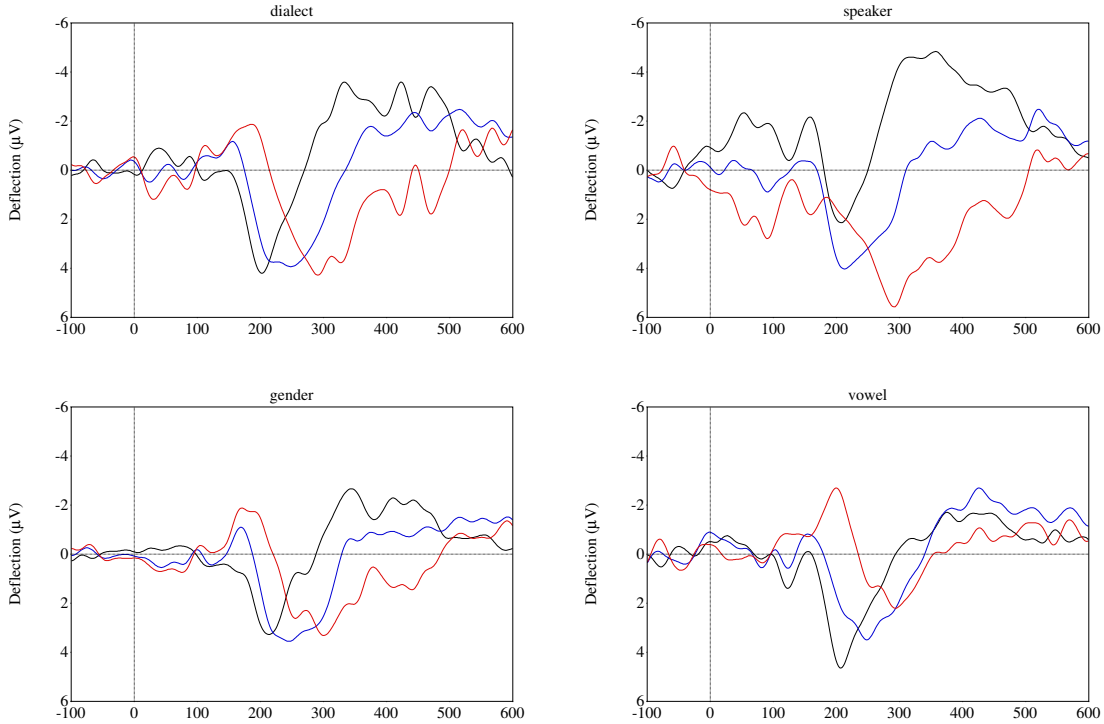

Figure 5. Grand-average waveforms at channel Fz for the control(black line), deviant stimuli(blue line) and difference waves(red line) per deviant type.

As can be observed in figure 6, the results from the current study point to the fact that gender change appears to yield an MMN amplitude comparable to vowel change indicating that listeners do not automatically normalize gender differences. Nevertheless, during the visual analysis very little differences were noticed between the two types of deviants (i.e. gender and vowel). The MMN amplitudes of dialect appear to be larger than MMN amplitudes elicited from change in speaker suggesting that listeners normalize differences in speaker more readily than changes in the dialect.

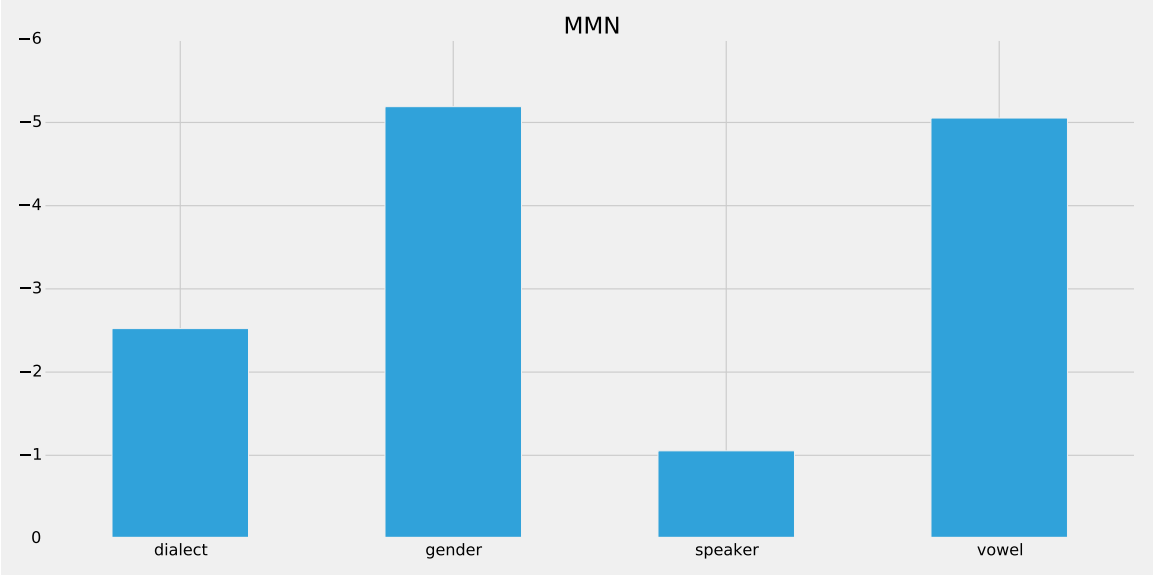

Figure 6. MMN at Fz channel for all deviant types

P3a components were elicited as well from ERP data. See figure 7. The results retrieved from the 40 ms time window from the channel Fz reveal that dialect and speaker deviants yielded a similar P3 response, speaker showing the highest value. Gender elicited a larger P3a effect than vowel category.

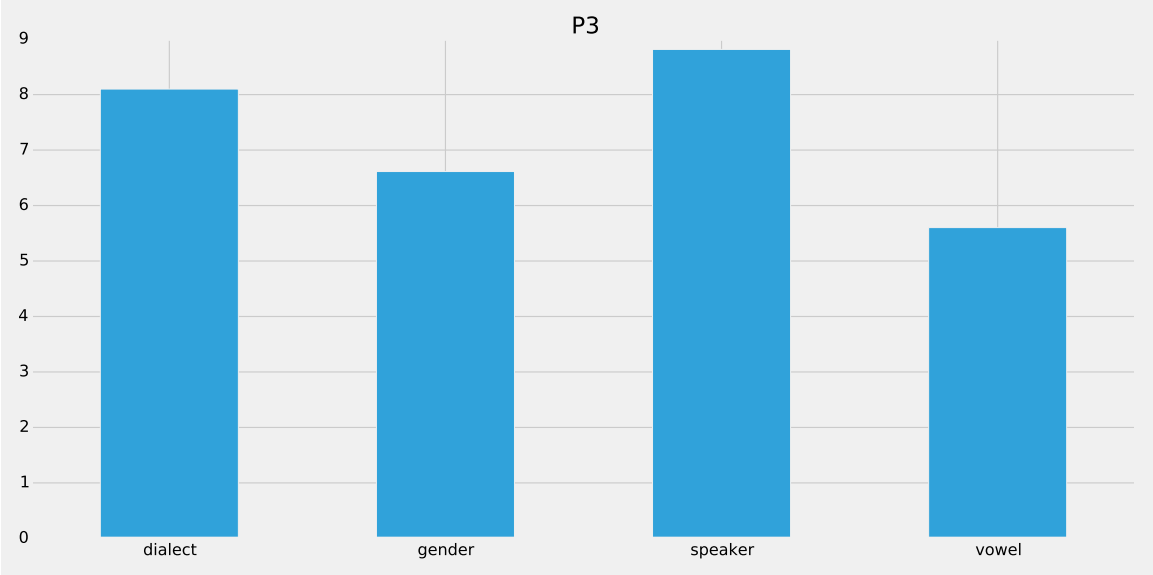

Figure 7. P3a at Fz channel for all deviant types

Discussion

The qualitative analysis of the present study seems to indicate that gender changes yielded larger MMN than vowel changes implying that listeners normalize changes in vowel more readily than changes in gender. Furthermore, dialect variation elicited a larger MMN than speaker variation. The weak mismatch responses to speaker variation indicate that listeners normalize differences in speaker identity and that they normalize speaker variation more readily than dialect variation. Result differences between the preceding experiments and the present study could be due to the modification of the deviant stimuli which in both of the previous studies differed in F0 to various extents from the standard and in the present experiment F0 differences from the standard were all approximately the same extent. As in the previous experiments on Australian English listeners and Dutch listeners, P3a differences from the current study indicate a combined effect of acoustic and linguistic processing. Inasmuch as P3a is an ERP component associated with attention and succeeding memory processing, listeners noticed all types of deviation (e.g. deviations in voice quality and vowel category). As predicted and in line with the second hypothesis, dialect change yielded a large mismatch response indicating that variation across dialects is not easily normalized by listeners and requires lexical or metalinguistic knowledge in order for the vowels to be fully recognized.

Very interestingly, even after the F0 differences were equalized listeners still notice to a great extent gender differences. This could indicate the fact that apart from the F0's important role in the perception of vowels, several differences such as the vocal tract length which has consequences on the production of the tone at the glottis, had a similar importance. More specifically, depending on the vocal tract configurations and the different frequency components that are strengthened, listeners hear different vowel qualities (Simpson, 2009). Therefore, the parameters involved in defining the formant pattern are speaker specific thus the formant frequencies of vowels produced by men and women are specific to each gender. The fact that the values of the formant frequencies depend on the length of the vocal tract of the speaker and are on average higher for females than for males,

could have an effect on the ability of listeners to perceive changes in gender as they could associate low or high formant frequencies to men, respectivly, women and therefore make a clear distinction between both genders.

Another possible effect could also be caused by the articulatory speed. Studies indicate that the differences in the average articulatory dimensions of males and females can have an effect on the average size of the acoustic vowel space, more specifically a larger female acoustic vowel space, which might trigger the perception of a faster speaking rate (Simpson, 2002). Therefore, females may be perceived as speaking at a faster tempo due to the fact that on average they cross a larger acoustic vowel space during the same time-frame than male speakers do (Weirich & Simpson, 2013). If males and females had a similar articulation speed, when articulating a vowel within the same time-frame, the acoustic realization of the vowels would be different because females would need a shorter amount of time to reach the vowel target in comparison to males who would need a longer amount of time to produce the same vowel. This would also imply that the articulatory distances between vowel categories for women would be acoustically more distanced than for men. Consequently, differences in the articulatory speed could result in different formant frequencies for the same vowels produced by males than those produced by females. As vowel formants play an important role in the perception of gender for listeners, such differences in the articulatory speed having as a result different formant values, would affect the perception of gender changes.

Therefore, in the present study as well as in the previous ones, listeners could have noticed the difference between vowels produced at a higher speaking rate and vowels produced at a lower speaking rate (i.e. vowels produced by women versus vowels produced by men) differently and thus, not normalize the differences in gender. In contrast to this, if the acoustic output of vowels produced by males and those produced by females were similar, this would entail that women are able to produce short vowel durations while increasing the articulatory speed.

The overall results obtained in the present study indicate that listeners are able to

detect gender variation in a pre–attentive task and that they normalize speaker differences more readily than all the other types of variation tested in the experiment. Future research with a larger number of participants should be carried out in order to confirm the results from the ERP present study.

References

Adank, P., Van Hout, R., & Van de Velde, H. (2007). *An acoustic description of the vowels of northern and southern standard dutch ii: Regional varieties* (Vol. 121:1130-41). Journal of the Acoustical Society of America.

Ainsworth, W. (1972). *Duration as a cue in the recognition of synthetic vowels* (Vol. 51:648–651). J. Acoust. Soc. Am.

Bennett, D. (1968). *Spectral form and duration as cues in the recognition of english and german vowels* (Vol. 11:65–85). Lang. Speech.

Berger, H. (1929). *Uber das elektrekephalogramm des menschen* (Vol. 87:527-70). Arch Psychiatr Nevenkr.

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer [computer program] version 5.4.12.* Retrieved from `http://www.praat.org/`

Bolt, R., Cooper, F., David, E., Denes, P., Pickett, J., & Stevens, K. (1970). *Speakers identification by speech spectrograms: A scientists' view of its reliability for legal purposes* (Vol. 47) (No. 2. pp:597-612). Journal of Acoustical Society of America.

Chládková, K., Boersma, P., & Podlipský, V. (2009). *On-line formant shifting as a function of f0.* Amsterdam Center for Language and Communication, University of Amsterdam, The Netherlands.

Chládková, K., Geambasu, A., Dadwani, R., Peter, V., Schiller, N., & Escudero, P. (in preparation). *Speaker and dialect variation are handled differently: behavioural and pre-attentive evidence.* Amsterdam Center for Language and Communication, University of Amsterdam, The Netherlands.

Collins, B., & Mees, I. (2003). *The phonetics of English and Dutch* (Fifth ed., Vol. 5: 25-36). Brill, Leiden, The Netherlands.

Cutler, A., & Blumstein, S. (2003). *International encyclopedia of linguistics. speech perception* (Vol. 4:154-158). Oxford University Press.

Dadwani, R., Peter, V., Chladkova, K., Geambasu, A., & Escudero, P. (2015). *Adult listeners' processing of indexical versus linguistic differences in a pre-attentive dis-*

crimination paradigm. Conference: International Congress of Phonetic Sciences-2015, At Glasgow-Scotland.

Fry, D. B., Abramson, A., Eismas, P. D., & Liberman, A. M. (1962). *The identification and discrimination of synthetic vowels* (Vol. 5:171-189). Language and speech.

Fujisaki, H., & Kawashima, T. (1968). *The roles of pitch and higher formants in the perception of vowels* (Vols. AU-16,73-77). IEE Transactions on Audio and Electroacoustics.

Garrido, M., Kilner, J., Stephan, K., & Friston, K. (2009). *The mismatch negativity: A review of underlying mechanisms* (Vol. 120:453-463). Clin. Neurophysiol.

Gottfried, T., & Chew, S. (1986). *Intelligibility of vowels sung by a countertenor* (Vol. 79:124-130). J. Acoust. Soc. Am.

Hallé, P., & Boysson-Bardies. (1994). *Emergence of an early receptive lexicon: Infants' recognition of words* (Vol. 17:119-129). Infant Behavior & Development.

Hunter, J. D., & Team, M. D. (2015). *Matplotlib.* Retrieved from `http://matplotlib.org/`

Kappenman, E., & Luck, S. (2011). *The oxford handbook of event-related potential components* (Vol. pp.4-7). Oxford Library of Psychology.

Krauss, R. M., & Pardo, J. S. (2006). *Speaker perception and social behaviour: Bridging social psychology and speech science* (Vol. pp. 273-278). P.A.M. van Lange (Ed), Bridging Social Psychology: The Benefits of Transdisciplinary Approaches.

Kriengwatana, B., Escudero, P., & Terry, J. (2014). *Listeners cope with speaker and accent variation differently: Evidence from the go/no-go task.* Poc. SST, Christchurch 2014.

Krishna, S., Patil, K., & Elhilali, M. (2012). *Recognizing the message and the messenger: biomimetic spectral analysis for robust speech and speaker recognition.* Int J Speech Techno.

Ladefoged, P., & Broadbent, D. (1957). *Information conveyed by vowels* (Vol. 29:29:98-104). Journal of Acoustical Society of America.

Lambda Foundry, I., & Team, P. D. (2015). *Pandas.* Retrieved from `http://pandas.pydata.org/`

Lehiste, I., & Meltzer, D. (1973). *Vowel and speaker identification in natural synthetic speech* (Vol. 16:356-64). Language and Speech.

Luck, S. (2014). *An introduction to the event-related potential technique* (Second ed.). Cambridge, MA: MIT Press.

Näätanen, R. (1992). *Attention and brain function.* Hillsdale,Nl: Lawrence Erbaum Associates.

O'Connor, K. (2015). *Vowels, vowel formants and vowel modification.* Retrieved from `http://www.singwise.com/cgi-bin/main.pl?section=articles&doc=VowelsFormantsAndModificat`

Paavilainen, P. (2013). *The mismatch negativity (mmn) component of the auditory event-related potential to violations of abstract regularities: A review* (Vol. 88) (No. 2. pp:109-123). International Journal of Psychophysiology.

Picton, W. T. (1992). *The p300 wave of the human event-related potential* (Vol. 9(4):456-479). Journal of Clinical Neurophysiology.

Polich, J. (2007). *Updating p300: An integrative theory of p3a and p3b* (Vol. 118(10):2128-2148). Clin. Neurophysiol.

Rosner, B. S., & Pickering, J. B. (2008). *Vowel perception and production.* Oxford Scholarship Online.

Simpson, A. (2002). *Gender-specific articulatory-acoustic relations in vowel sequences* (Vol. 30: 417–435). Journal of Phonetics.

Simpson, A. (2009). *Phonetic differences between male and female speech* (Vol. 3/2:621–640). Language and Linguistics Compas.

Slawson, A. W. (1968). *Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency* (Vol. 43:87-101). JASA.

Sundberg, J. (1977). *The acoustics of the singing voice* (Vol. 236:82). Scientific American.

Traunmüller, H. (1981). *Perceptual dimension of openness in vowels* (Vol. 69:1465-1475). JAS.

Weirich, M., & Simpson, A. (2013). *Acoustic vowel space size and perceived speech tempo*

(Vol. 19:5-10). Acoustical Society of America.

Wood, S. (2015). *Fft silices: Wideband and narrowband slices.* Retrieved from http://swphonetics.com/praat/snded/fftslices/