PROCESSING OF ACOUSTIC CUES FOR VOICING IN ENGLISH: A MMN STUDY

Outi Tuomainen[1], *Heather van der Lely*[1]

[1] UCL Centre for Developmental Language Disorders and Cognitive Neuroscience, UCL, London, o.tuomainen@ucl.ac.uk, h.vanderlely@ ucl.ac.uk

ABSTRACT

Speech perception normally utilizes multiple acoustic cues in perception of specific speech sound contrast. This study investigates which acoustic cues are responsible for syllable final stop consonant voicing in English using speech and non-speech stimuli. Specifically we study vocalic duration and F1 offset frequency cues using three experimental paradigms. Two paradigms used behavioural methods and explored identification (Exp1) and discrimination (Exp2) and one an electrophysiological method to investigate the neural correlates of processing in a mismatch negativity (MMN) experiment (Exp3). In Exp1 we presented the [bot]-[bod] continuum varying either in duration or F1 cues. Exps 2 and 3 employed a 2 (Frequency: high low) x 2 (Duration (long, short) design resulting in four different versions of English non-words [bot] and [bod] and their corresponding non-speech analogues. Nine subjects participated in Exp 1 and eight in Exps 2 & 3. The findings from Exp 1 revealed that the duration cue plays an important role in British English syllable final stop voicing. Further support for this finding was revealed in Exp 3 with larger MMN amplitude for the duration cue compared with the frequency cue.

Keywords: speech perception, cue-weighting, Event-related potentials, mismatch negativity (MMN)

1. INTRODUCTION

It is generally accepted that different speech-sound contrasts are cued by multiple different parameters of the acoustic signal. The listener, however, does not treat all these cues as perceptually equivalent and may assign more weight to some of these cues over others [8]. Interestingly, this tendency is not consistent across development[6] but appears to be linked with phonemic development and awareness. Better phonemic awareness affects the cue weighting strategies [4]. Moreover, even when the cues are equally informative and discriminable, people weight cues differently in categorization tasks [3]. In this study, we investigated the relative weight assigned to acoustic cues that signal syllable final consonant voicing in English; namely vocalic duration and first formant (F1) offset frequency. We investigated how these cues are discriminated both attentively and preattentively and whether or not prototypicality affects performance. We use a similar cue-weighting design and stimuli as previously employed by Crowther & Mann ([1]) and Nittrouer ([5]). In addition to an identification task, we administered two discrimination experiments: a behavioral discrimination task and an MMN experiment.

Based on previous work in perception and production ([2], [5]), we predict that the vocalic duration will be strongly weighted by adult British English speakers. Moreover, we predict a preference of duration cue over F1 offset cue in the discrimination tasks; both in the attentive Exp2 and pre-attentive Exp 3, indexed by the mismatch negativity component (MMN) of auditory event-related potentials (ERPs, for a review, see [7]). In addition to comparing the identification and discrimination performance, our design allows us to investigate if prototypical stimuli are processed differently from non-prototypical stimuli.

2. METHODS

2.1. Participants

Nine healthy monolingual native British English adult volunteers participated in the identification task eight of which (mean 23;6 years, range 17;11-36;5 3 men) took part in the EEG recording. We recruited participants via student email lists and they were paid £10 for participation. All subjects were right handed (the Edinburgh handedness questionnaire) and reported normal hearing and no known neurological condition or history of language impairment.

2.2. Stimuli and procedure

2.2.1. Behavioral tasks

There were two behavioral tasks in the present experiment: an identification task (Exp1) and a discrimination task (Exp2). In Exp1 two [bod-bot] continua of seven stimuli were synthesized (2x7 design). We varied the duration of the vocalic portion while keeping the frequency of the first formant constant (F1 frequency is either "high" as in voiceless consonants, continuum 1, or "low", as in voiced consonants, continuum 2), consistent with previous work investigating cue-weighting in American English [1], [5]. The stimuli were synthesized with High-Level Speech Synthesizer (HLsyn, Sensimetrics Inc., 1.0). For the "high" continuum, the F1 frequency was 570 Hz throughout the stimuli and the vocalic portion varied from 100 ms to 220 ms in 20 ms steps. For the "low" continuum, the F1 offset frequency was lowered to 250 Hz during the final 50 ms of the vocalic portion. The vocalic duration varied from 100-220 ms in 20 ms steps. (For a thorough description of the synthesis, see [1].) A 50 ms period of silence signaling an initial stop consonant proceeded each vocalic portion.

In Exp 1 all 14 stimuli from the two continua were played 10 times at a pseudo-random order, one sound at a time. The subjects were asked to identify the stimulus as the English non-words [bot] or [bod] by pressing a key on the keyboard. A short practice session (15 stimuli, in a fixed order) preceded the experiment to establish that subjects heard the stimuli as [bot] and [bod]. The identification task took approximately 5-7 minutes to complete.

For the behavioral discrimination tasks four stimuli (from the 14 synthesized) were chosen based on the 2 (Duration: long, short) x 2 (Frequency: high, low) design. Thus, two of the stimuli were reliably identified as [bot] (vocalic duration short 120 ms, F1 high, 570 Hz) and [bod] (vocalic duration long 220 ms, F1 low 250 Hz) and possibly represented the prototypical exemplars of syllable-final British English [t] and [d] (prototypical stimuli are named as "high120" and "low220"). The other two stimuli in Exp 1 were reliably identified as either [bot] or [bod] but they contained conflicting cues for the consonant in question. In other words, they consisted of formant transition typical for voiceless consonant but vocalic duration typical for voiced one (named "high220" identified as [bod]) and vice versa (named "low120" identified as [bot]) thus forming non-prototypical within-category variants of the prototypical non-words. To keep the behavioral and preattentive (MMN) paradigms identical, the stimuli were presented in a roving-standard paradigm [8] where each deviant becomes the standard stimulus (SOA 1000 ms, 187 stimuli in total, 40 deviants). In this paradigm, the number of standards preceding a deviant was random, i.e. the change was not predictable. For Exp 1, we presented stimuli via headphones with a lap top computer in a quiet room. In Exp 2 subjects were asked to press a button as quick as possible as soon as they heard a change in the stimulus train (go/no-go task). A short practice session (total of 34 stimuli, 5 deviants) preceded the experiment to make sure subjects understood the instructions. The experiment took approximately five minutes to complete.

In addition to speech stimuli described above for Exps 2 and 3 non-speech analogues of the speech stimuli were created by replacing the F1-F3 tracks with sinusoids (Praat, 4.4.1.6). Thus, this created a relatively close match of the speech signal in a non-speech control sequence. Subjects were told that these non-speech sounds were synthetic noise or science fiction sounds. Presentation of the speech and non-speech sounds was identical in Exps 2 and 3. Subjects described the non-speech stimuli after the experiment. None of the subjects reported hearing the sounds as speech. We balanced the order of presentation of the speech and non-speech conditions across subjects.

2.2.2. EEG recording

The same stimuli (speech and non-speech) and the same paradigm (roving-standard, SOA 800 ms, total 2160 stimuli) as in the behavioral discrimination tasks were used in the EEG recordings. There were 120 deviants in each category (total 480 deviants). Those standards immediately following a deviant were removed from analysis. EEG was recorded with 128 channel electrode net (Electrical Geodesics Inc.) using Net Station (4.1.2) software for data acquisition and analysis. Amplifier sampling rate was 250 Hz with a 0.1-100 Hz band pass filter. The stimuli were presented with Biological E-Prime Program (Psychology Software Tools, Inc. version 1.0.20.1) via loudspeakers at a comfortable level while subjects were seated in a Faraday cage in a comfortable chair. Subjects watched children's

cartoons, conducted a simple counting task, and were asked to ignore the auditory stimuli. Subjects' performance in the counting task was monitored. Exps 2 and 3 consisted of two separate sessions (speech and non-speech) each of which was divided into four nine minute blocks with a short break between the blocks. The entire recording session took approximately 90 minutes. EEG data were off-line filtered with a 1-30 Hz band pass filter, baseline corrected with respect to 100 ms prestimulus baseline. A 70 µV artifact detection criterion was used. The data were segmented from 100 ms prestimulus to 600 ms poststimulus and averaged offline. Finally the data were re-referenced to the average voltage.

3. RESULTS

3.1. Exp 1: Identification

In the identification data, the ratio of [bod] responses to each stimulus was calculated (see Figure 1). The identification data from synthetic [bot] [bod] continua were analyzed with probit regression analysis that gives an estimation of the Point of Subjective Equivalence (PSE) and the slope of the categorization function. These data were then analyzed with 2x7 ANOVA and paired sample t-tests.

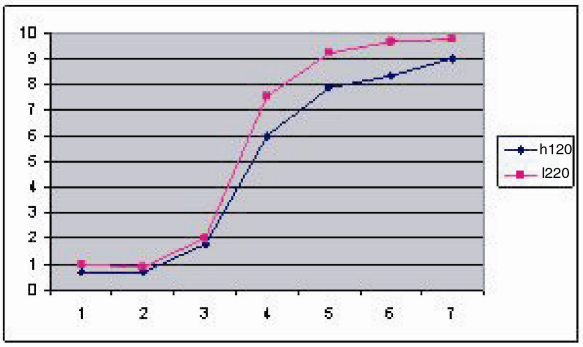

Figure 1: The labeling function of synthetic [bot] - [bod] continua from nine subjects (number of [bod] responses out of 10). Label "bot" refers to "high" offset and "bod" to "low" offset.

Probit analysis showed that the category boundary differs significantly in these two continua ([bot] continuum M=4.1, sd = 0.35; [bod] continuum M=3.4, sd = 0.47; (t(8)=2.669, p=0.028)). The difference in the slope of these two continua approached significance (t(8)=-2.277, p=0.052).
The main effect of continuum (i.e. the formant offset frequency) was significant (F(1,8)=8.244, p=0.021) indicating that these two continua differ overall. However the interaction stimulus x continuum was not significant (F(6,48)=1.383, p=0.241).
These data suggest that duration is a prominent cue for syllable-final consonant voicing in British English as expected on the basis of previous results of [2], [5]. This is relatively clear at least within category stimuli (see Figure 1). However, the formant offset information plays a role in identifying the voiced-voiceless consonants (as indexed by the slope and PSE). In other words, when the vocalic portion is short/long the formant transition plays a less important role in identification, whereas when the durational cue is more ambiguous, the listener begins to use the formant offset cue as well.

3.2. Exp 2: Discrimination

In the discrimination task the sensitivity to cues was quantified by calculating the d' for each subject and condition (see Table 1).

Table 1. Sensitivity (d') for different experimental conditions (sd in parenthesis).

	High120	High220	Low120	Low220
Speech	1.59 (0.35)	1.58 (0.30)	1.33 (0.39)	1.52 (0.16)
Non-Speech	1.37 (0.49)	1.27 (0.42)	1.00 (0.40)	1.10 (0.43)

2x2x2 Repeated measures ANOVA (Mode: speech vs. non-speech; Duration: short vs. long; Formant: high vs. low) revealed a main effect of frequency (F(1,7)=12.688, p=0.009) which was due to higher sensitivity in the "high" formant transition condition. The main effect of mode approached significance (F(1,7)=5.052, p=0.059). In general, the speech stimuli were discriminated slightly better than non-speech stimuli. No other significant main effects or interactions were found.

3.3. Exp 3: MMN for speech and non-speech stimuli

All amplitude measurements and statistical analyses were conducted on the central electrode Cz. Mean MMN amplitude was computed from the grand averaged difference waveforms in a fixed time-window of 50 ms (295-345 ms after stimulus onset, i.e. the maximum peak occurring approximately 200 ms after the deviation point). Statistical analyses were performed in a 2x2 repeated measures ANOVA. The difference waves are created by subtracting the standard from the

deviant within the same sound category. The group grand average mean amplitudes (in μV) are presented in Table 1 and the corresponding MMN grand average difference waves in Figure 2.

Table 2. The mean amplitudes (standard deviation) of speech and non-speech stimuli.

	High120	High220	Low120	Low220
speech	-0.78 (1.3)	-0.72 (1.3)	-1.45 (1.1)	-0.03 (1.4)
non-speech	-0.65 (1.5)	-0.11 (1.0)	-1.32 (1.1)	-0.58 (0.8)

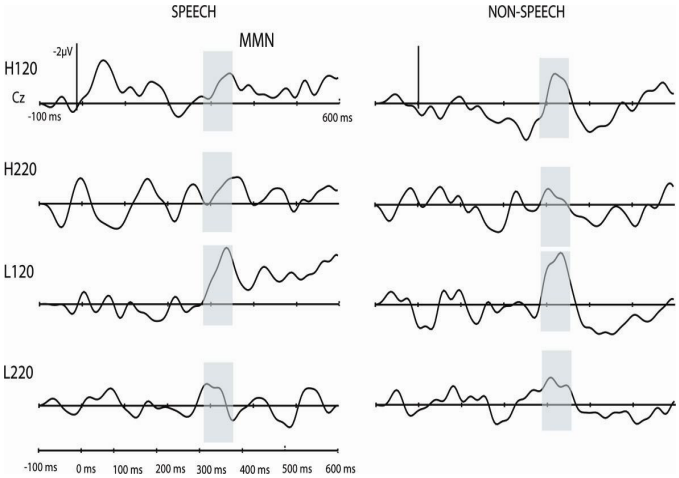

Figure 2. The Grand Average difference waves for speech and non-speech stimuli from six subjects.

Statistical analyses revealed no differences between speech and non-speech stimuli ($F(1,7)=.062$, $p=.811$). However, the duration cue (short vs. long) was significant ($F(1,7)=8.883$, $p=0.021$) whereas the F1 offset frequency (low vs. high) was not ($F(1,7)=.793$, $p=0.403$). The formant x duration interaction indicating prototypicality did not reach significance ($F(1,7)=2.402$, $p= .141$).

4. CONCLUSIONS

The behavioral findings indicate that both frequency and duration cues are used in identifying English /t-d/ contrast. However, the identification experiment indicated that the durational cue has a prominent role in within category stimuli perception. These results concur with previous studies on weighting of acoustic cues in syllable final voicing in English. The brain's automatic change-detection (MMN) revealed a similar pattern to the identification data. This was revealed by a larger MMN amplitude for the durational cue than for the frequency cue. However, the discrimination data from Exp 3 indicated increased sensitivity to frequency deviants. This discrepancy between behavioral discrimination sensitivity and MMN could be due to attentional effects. However, the relatively low d' in Exp 2 indicates that this task was difficult and could have resulted in subjects' using strategies which could account for the different Exp2 and 3. Alternatively, it could indicate qualitative differences depending on the stage of processing.

In contrast, we found that the prototypicality of the stimulus did not have an effect on behavioral discrimination accuracy or the MMN amplitude. Moreover, the speech and non-speech conditions did not differ in MMN amplitude. This could indicate that the durational cue is a perceptually salient cue in general auditory processing at the auditory cortex. Alternatively, it could indicate that once attention to a specific cue has developed in speech perception this cue is always attended to (auditory pop-out effect) regardless of the stimuli status. Further investigations are warranted to tease these possible interpretations apart.

5. REFERENCES

[1] Crowther, C.S. & Mann, V. 1992. Native language factors affecting use of vocalic cues to final consonant voicing in English. *J.Acoust.Soc.Am.* 92, 711-722.

[2] Flege, J. et al. 1992. Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin and Spanish. *J.Acoust.Soc.Am.* 92, 128-143.

[3] Holt, L. & Lotto, A.J. 2006. Cue weighting in auditory categorization: Implications for 1st and 2nd language acquisition. *J.Acoust.Soc.Am.* 119, 3059-3071.

[4] Mayo, C. et al. 2003. The influence of phonemic awareness development on acoustic cue weighting strategies in children's speech perception. *JSHR*, 46, 1184-1196.

[5] Nittrouer, S. 2004. The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. . *J.Acoust.Soc.Am.* 115, 1777-1790.

[6] Nittrouer, S. & Miller, M.E. 1997. Predicting developmental shifts in perceptual weighting schemes. *J.Acoust.Soc.Am.* 101, 2253-2266.

[7] Näätänen, R. 2001. The perception of speech sounds by the human brain as reflected by mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, 38, 1-21.

[8] Ohde, R.N. &Haley, K.L. 1997. Stop-consonant and vowel perception in 3- and 4-year-old children. *J.Acoust.Soc.Am.* 102, 3711-3722.

[9] Shestakova, A. et al. 2002. Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. *Neuroreport*, 13, 1813-1816.