Spontaneous and explicit speech imitation

Jeesun Kim, Ruben Demirdjian, Chris Davis

The MARCS Institute, University of Western Sydney, Australia

j.kim@uwsedu.au, reuben.demirdjian@gmail.com, chris.davis@uws.edu.au

10.21437/Interspeech.2013-152

Abstract

The single word speech shadowing task typically produces spontaneous imitation of the speech of the shadowed talker (the model). This task has been used as a tool for examining speech convergence in a non-social setting and has provided data for claims that the mental lexicon is constituted from instance-based exemplars. We examined whether the speech of participants who shadowed or explicitly imitated a model talker would converge on similar properties of the model's speech. The model talker produced speech in two styles (normal and clear speech) that differed in word duration, intensity and F0. Participants produced immediate and delayed naming responses. The results suggested that spontaneous and explicit imitation tap different processes. For spontaneous imitation only word duration showed convergence with model speech and this effect was reduced with delayed naming. Explicit imitation showed an association with model speech both for duration and intensity and this effect was unaffected by the delayed naming. The pattern of partial correlations between the imitation conditions and the model speech provided further evidence that the spontaneous imitation was based upon different processes than those used in explicit imitation.

Index Terms: Speech Shadowing, Spontaneous imitation, Explicit imitation

1. Introduction

In the speech shadowing task [1] a participant repeats a word as soon as she/he can. Marlen-Wilsen [2] has described this procedure as a classical "black-box" paradigm in which the transduction of speech input to output provides an index of the system's transfer-function. In a now classic paper, Goldinger [3] showed that participants in a single word shadowing task exhibited a tendency to spontaneously imitate the shadowed word and he interpreted this result as suggesting that the mental lexicon is constituted by detailed episodic instances. In its ability to elicit spontaneous imitation, the shadowing paradigm also provides a well-controlled method to investigate speech convergence in a non-social setting. However, Goldinger suggested that there was an interpretative problem in his study in that "imitation data are only theoretically relevant if they reflect a spontaneous response from memory to spoken words" and that "…listeners may have a frivolous tendency to imitate voices, regardless of deeper lexical processes" (p.256). The current experiment aimed to investigate this issue by contrasting the results of the shadowing paradigm with those produced by explicitly instructing participants to imitate the model talker.

In order to gauge imitation performance Goldinger [3] used an AXB procedure in which listeners heard a shadowed token (A or B) and a read aloud baseline token (B or A) of the same word and judged whether A or B sounded a better imitation of the model X. This perceptual assessment provides a global index of imitation but it does not provide information about which aspects of the speech signal contribute to the assessment of imitation. In this regard, Goldinger [3] suggested that word duration, amplitude and fundamental frequency (F0) might provide useful measures to index imitation. Indeed, in an informal test of which of these properties might be important, Goldinger selected items where AXB classification performance yielded high rates (92%) of selecting the shadowed utterance as a match to the model and then modified the duration, amplitude or F0 of these tokens. He found that equating duration (but not amplitude or F0) significantly decreased the detection of imitation. Based on this our examination focused on measuring word duration, amplitude and F0 with the expectation that duration may be the one that converges most.

In order to detect imitation of the Model's speech we had the Model talker produce speech in two styles, Normal and Clear speech since tokens produced in these styles should differ in duration, amplitude and F0. We also contrasted having imitators immediately utter a response or utter it only after a delay. This manipulation was included because Goldinger found that delayed naming reduced the impact of spontaneous imitation; it would thus be useful to assess the impact of this manipulation on explicit imitation.

2. Method

2.1. Participants

There were eighteen participants (9 females, aged 30-45; males 32-47). All were native speakers of Australian English and were paid $15 for participation.

2.2. Materials

Seventy-two bi-syllabic, low-frequency English words were used as stimuli. These words were selected from the list used by Shockley et al. [4]. These stimuli were selected to have low frequency of occurrence (M = 11.4 occurrences per million; SD = 10.0 [5]) since low frequency words have been shown to elicit a higher degree of imitation [3]. All stimuli began with the voiceless stop consonants (/k/, /p/ or /t/) to ultimately enable a straightforward determination of VOT. These 72 words were split into four groups of 18 words. These 18 words consisted of six words each beginning with the consonants /k/, /p/ and /t/ and were matched in terms of frequency.

2.3. Recording Procedure

To generate the spoken versions, the head and face (from shoulders up) of a male speaker (the model speaker, Australian native speaker, 25 years old) were recorded (using a Sony HVR-VIP and AKG C417 lavalier microphone) against a dark green background and were illuminated with a key and fill lights (three Photon Beard 110 studio lights).

In the recording session, each word was presented singly to the model speaker on a computer monitor (placed directly below the camera). After reading the word, the model speaker said it aloud in a natural speaking voice (Normal Speech). There was a gap of five seconds between each word to be read aloud. Following this session, the model speaker read the words a loud again but this time articulated each carefully and clearly (Clear Speech). In both sessions several tokens were recorded and one was selected for the test session based upon clarity of pronunciation, sound quality and the naturalness of the facial movements.

Eight additional words were also recorded as practice items. The video files were edited so that the beginning and end of each token showed the model speaker with a neutral expression and closed mouth.

2.4. Test Procedure

Participants were tested individually in an IAC booth. Stimuli were presented using the DMDX software [6] with audio presented via a Beyerdynamic DT-297 headset and participant responses recorded via the headset's condenser microphone.

As a Baseline condition, participants first read aloud all 72 words that were presented one at a time on the monitor (presentation order was randomized over participants). The design and condition presentation order for the spontaneous imitation (shadowing) and explicit imitation conditions that were presented after the Baseline condition is shown in Table 1. Following the Baseline condition, participants first completed the Spontaneous imitation conditions (Normal and Clear speech) and then the Explicit imitation conditions (Normal and Clear speech). The Normal and Clear speech conditions (presented within each of the imitation conditions) were blocked (36 items in each) and the order of block presentation was counterbalanced across participants. Participants were not informed when the articulation style changed. The Immediate and Delayed conditions (18 items each) were presented blocked with the Immediate items presented first.

Table 1. *Design of the imitation tasks, articulation style and response delay condition.*

	Spontaneous Imitation		Explicit Imitation	
	Immediate	Delayed	Immediate	Delayed
Normal speech	Words 1-18	Words 19-36	Words 1-18	Words 19-36
Clear Speech	Words 37-54	Words 55-72	Words 37-54	Words 55-72

In the Spontaneous imitation conditions each participant was told that she/he should say aloud the heard word into the microphone quickly but clearly. In the Explicit imitation condition, participants were told to try to imitate or mimic the sound of the talker's voice. Participants were asked to respond either immediately on hearing the word (Immediate response condition) or after a two second delay in which "babble speech" was presented (Delayed condition). To ensure that the delay instruction was adhered to, a picture of a microphone was presented as a cue to vocalize across conditions.

2.5. Data Processing

The duration, intensity and F0 of the recorded words were analyzed using Praat [7].

3. Results

The first issue to address is whether there was imitation (spontaneous or explicit) for any of the measured properties (duration, amplitude, F0). This was examined by measuring a change for the model speaker for the normal versus clear speech contrast and comparing this to any change produced by the participants. The rationale was that the difference in the model speaker's utterances would provide the basic variation over which imitation performance could then be determined. The difference between the duration of the clear and normal speech tokens for the Baseline, imitation conditions and the Model talker is shown in Figure 1.

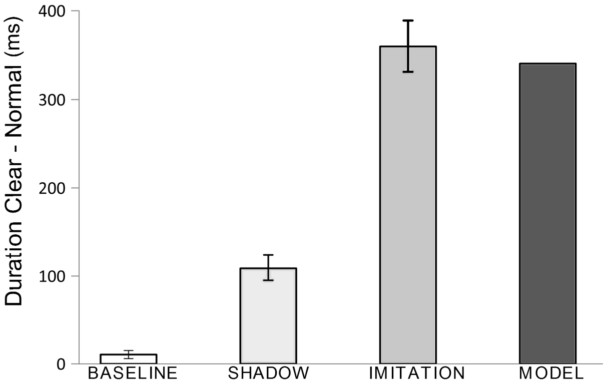

Figure 1: Average *Clear speech minus Normal speech durations (ms) for the immediate response condition. Note SHADOW = Spontaneous Imitation (error bars = SEM).*

The duration data for the baseline and imitations conditions (Immediate conditions) were compared in a series of repeated ANOVAs using the participant data and using a non-repeated ANOVA on the item data for comparing the Normal and Clear speech of the model talker (Bonferroni corrected for multiple comparisons). There was a difference between the duration of normal and clear speech for the model talker, F(1,71) = 96.94, $p < 0.05$, ŋp2 = 0.58 (a clear speech effect). To determine whether this difference was shown in the Spontaneous imitation condition, the difference between Clear and Normal speech durations for this condition was compared to the same contrast in the Baseline condition. There was a significant difference between Baseline vs. Spontaneous Imitation condition, F(1,17) = 26.69, $p < 0.05$, ŋp2 = 0.60. There was also a significant difference between the duration of the Spontaneous Imitation and the Explicit imitation condition, F(1,17) = 85.58, $p < 0.05$, ŋp2 = 0.60. There was no difference between the size of the clear speech effect for the Explicit imitation condition and that found for the model talker, $F < 1$.

Two ANOVAs were conducted to determine the effect of delayed naming on the two imitation conditions. For the Spontaneous Imitation condition, the clear speech effect was 107 ms in the immediate response condition and was reduced to 68 ms with delayed naming. This reduction was significant with delayed naming responses being shorter (less influenced by the longer clear speech token, F (1,17) = 7.68, $p < 0.05$, ŋp2 = 0.13. For the Explicit imitation condition, the 370 ms clear speech effect was only reduced to 350 ms in the delayed naming condition and this difference was not significant, F(1,17) = 1.64, p > 0.05.

The average difference between the amplitude (dB) of the Clear and Normal speech tokens for the baseline, imitation and model talker conditions is shown in Figure 2.

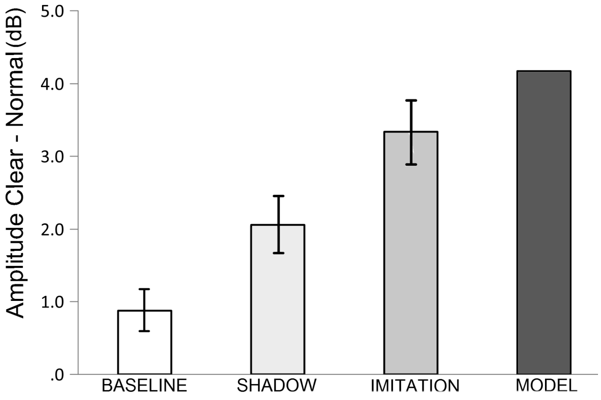

Figure 2: Average *Clear speech minus Normal speech amplitude (dB) for the immediate response condition. Note SHADOW = Spontaneous Imitation (error bars = SEM).*

For the Immediate response condition, there was a significant effect of clear speech on amplitudes for the Model talker, with clear speech having a greater amplitude (55.9 dB) than normal (51.32 dB) speech (non-repeated ANOVA item data), F(1,71) = 44.23, p < 0.05, ŋp2 = 0.39.

This difference was not found in the Spontaneous Imitation condition as a repeated ANOVA comparing the Baseline amplitude vs. the Spontaneous Imitation condition was not significant, F(1,17) = 1.71, p > 0.05. There was a clear speech amplitude effect for the Explicit imitation condition compared to the Baseline, F(1,17) = 10.42, p < 0.05, ŋp2 = 0.38. The size of this effect for Explicit imitation did not differ from that found for the Model talker, F(1,17) = 1.1, p > 0.05. Furthermore, the clear speech effect on amplitude for Explicit Imitation was not affected by delayed naming, with the difference between the Baseline and Explicit imitation condition still present in the delay condition, F(1,17) = 14.77, p < 0.05, ŋp2 = 0.47.

The difference between the F0 of the clear and normal speech tokens for all conditions is shown in Figure 3. For the Immediate response condition there was an F0 Difference for Model talker between the Normal (M = 172 Hz) and Clear (M = 203) speech styles, F(1,71) = 14.67, p < 0.05, ŋp2 = 0.17 (non-repeated ANOVA on the item data).

This raising of F0 shown by the model talker was not reflected in the imitation data as a difference between clear and normal speech styles, i.e., no F0 difference between the Spontaneous imitation condition and Baseline, F < 1 and no difference between the Explicit imitation and baseline, F < 1. Since there was a difference in the mean F0 of the female and male renditions (female M = 217 Hz, males M = 120 Hz, F(1,16) = 48.25, p < 0.05, ŋp2 = 0.97) the Explicit imitation scores of the female and male participants were considered separately. Explicit imitation of the Model by male participants led to a rise in F0 by 19.5 Hz in the Clear condition compared to the Baseline. The females participants had a slightly lower F0 in the Clear condition (-4.7 Hz). However, this difference between females and males was not significant, F(1,16) = 3.91, p = 0.065.

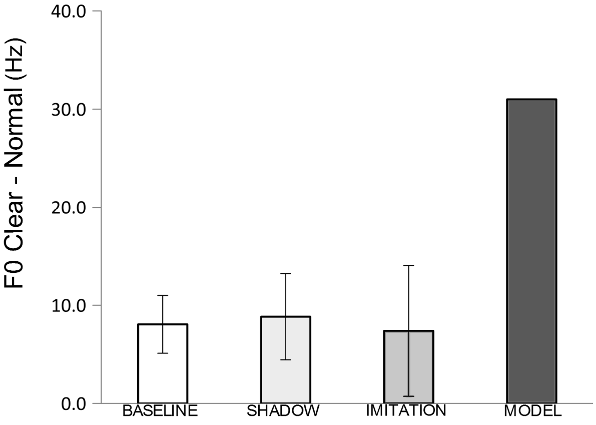

Figure 3: Average *Clear speech minus Normal speech F0 (Hz) for the immediate response condition. Note SHADOW = Spontaneous Imitation (error bars = SEM).*

The above data (e.g., the divergent effect of delayed naming) suggest that the processes that generate spontaneous and explicit imitation are different. It could be, however, that explicit imitation is simply more sensitive to changes in the speech signal and that the difference between the two imitation types was one of degree and not kind. If this were the case, it would be expected that Explicit and Spontaneous imitation would be correlated (i.e., pattern in a similar way as a function of variation in the properties of the Model talker's speech). What would not be expected on this sensitivity argument is that Spontaneous imitation would show a greater correlation with the Model talker than would Explicit imitation.

To examine the degree of association between Spontaneous and Explicit imitation with variation in the speech of the Model talker, a series of correlation analyses was conducted between the data of the two imitation conditions and that of the Model talker. These analyses were restricted to duration (since amplitude and F0 did not show an effect across the Clear and Normal speech manipulation for Spontaneous imitation) and the variation of the Baseline condition (due to word length, etc) was partialled out.

Durations in the Spontaneous imitation condition (Normal speech, Immediate response) showed an average partial correlation of 0.44 (p < 0.05) with those of the Model talker (Normal speech). When looked at on a per participant basis, it turned out that the partial correlations for 9 of the 18 participants were significant (p < 0.05). The Explicit imitation condition had an average partial correlation with the Model talker of 0.59 (p < 0.05). The partial correlations were secure for 12 out of the 18 participants. There was no correlation between the size of the partial correlations in the two imitation conditions (Pearson's r = 0.02, p > 0.05). Indeed, almost half of the participants who showed a significant partial correlation for Spontaneous imitation did not show a significant effect for Explicit imitation.

A similar result was found for the imitation duration data with the Model talker's Clear speech. For Spontaneous imitation (Clear speech, Immediate response) there was an average partial correlation with the Model talker's Clear speech duration of 0.51. The partial correlations of 11 participants were significant. The average partial correlation for the Explicit imitation condition with Model was 0.58 with the data from 12 participants showing significant partial correlations. The correlation between the size of the partial correlations in the two imitation conditions was not significant (Pearson's r = 0.18, p > 0.5).

Although Spontaneous imitation (as measured by the Clear vs. Normal speech contrast) was reduced in the delayed naming condition, the duration of shadowed Clear speech was still greater in this condition than in the Baseline one. Given this, we examined the correlation for duration (partialling out the Baseline data) between the Spontaneous imitation condition and the Model speech and between the Explicit imitation condition and the Model speech.

For Normal speech, the average partial correlation between the spontaneous speech and the Model talker's speech was 0.49 (p < 0.05). The partial correlations for 8 of the 18 participants were significant (p < 0.05). For Explicit imitation, the average partial correlation was 0.78 (p < 0.05) and the partial correlations were secure for all of the 18 participants. The correlation between the size of the partial correlations in the two imitation conditions was not significant (Pearson's r = 0.26, p > 0.05).

For Clear speech (Delayed naming), the average partial correlation between the Spontaneous imitations and the Model was 0.32 (p < 0.05). However, in the individual participant data only three participants showed a significant partial correlation. For the Explicit imitation condition (Clear speech, Delayed naming) the average partial correlation was 0.41 (p < 0.05) with the individual data from 10 participants showing significant partial correlations. The correlation between the size of the partial correlations in the two imitation conditions was not significant (Pearson's r = 0.1, p > 0.05).

4. Discussion

The current study examined spontaneous and explicit imitation performance by having a Model talker vary duration, amplitude and F0 by producing Normal and Clear speech. For Spontaneous imitation (shadowing) only the difference in duration had an influence and this influence was significantly reduced with delayed naming. For Explicit imitation, both the duration and amplitude changes of the model talker's speech were reflected in participants' productions and these effects were not reduced in the delayed naming condition.

The dissociation between Spontaneous and Explicit imitation was also supported by the distinct patterns of partial correlations between the word durations in the speech produced by each type of imitation and those of the Model. Such dissociation suggests that properties of the speech signal can be maintained by different memory systems, one supporting shadowing, and the other, explicit imitation.

In regard to the former system (the one tapped by shadowing), Goldinger proposed a form of lexical memory in which previously encountered spoken words are stored as a collection of memory traces. Retrieval from this system (what Goldinger calls "echo content") occurs when the newly presented word acts to probe these traces (which are activated in parallel to the degree that each resembles the probe). Output (echo content) consists of a combination of the newly presented probe and the activated traces. Such a system will reflect the idiosyncratic properties of the probe word to the degree that these are sparsely represented in lexical memory (i.e., under-represented properties will be less swamped by the aggregate echo). A memory system tapped by explicit imitation might be something like the episodic buffer proposed by Baddeley [8] as an extension of the three component working memory model of verbal STM [9].

Interestingly, a similar separation of memory systems has been proposed by Wise and colleagues [10]. Here a cut was made between a memory system specialized for the transient representation of sound sequences and a system that is specialized for the process of the mimicry of sounds. The former system (proposed to be located in the posterior left superior temporal sulcus) represents an externally presented word and interfaces with stored representations (i.e., the retrieved "internal" word). The latter system (located in the posterior-temporal/inferior-parietal junction) bridges the posterior temporal cortex and the motor speech system and as such allows for direct vocal mimicry of acoustic signals.

The above proposal of similar but different memory systems handling spoken input/output functions is similar to what Hickok and Poeppel [11] identified as the basic thesis of their approach: "that the execution of different linguistic functions involve non-identical neural networks, even with stimulus conditions held constant" (p.73). That is, when instructed to shadow, participants draw on a lexical system whereas when asked to imitate, participants monitor and use something like the episodic buffer of verbal STM.

5. Acknowledgements

The project was run as the second author's honours thesis. The authors thank Leo Chong and Michael Fitzpatrick for help with stimulus preparation and acknowledge support from Australian Research Council grants (DP0666857 & DP120104298).

6. References

[1] Chistovich, L.A., "Classification of rapidly repeated speech sounds", Akusticheskii Zhurnal, 6: 392-398, 1960.

[2] Marslen-Wilson, W.D., "Speech shadowing and speech comprehension", Speech Communication, 4: 55-73, 1985.

[3] Goldinger, S.D., "Echoes of echoes? An episodic theory of lexical access", Psychological review, 105: 251, 1998.

[4] Shockley, K., Sabadini, L., and Fowler, C.A., "Imitation in shadowing words", Attention,Perception, & Psychophysics, 66: 422-429, 2004.

[5] Kučera, H. and Francis, W. N., "Computational analysis of present-day American English" Dartmouth Publishing Group, 1967.

[6] Forster, K.I. and Forster, J.C., "DMDX: A Windows display program with millisecond accuracy", Behavior Research Methods, 35: 116-124, 2003.

[7] Boersma, P., "Praat, a system for doing phonetics by computer", Glot international, 5: 341-345, 2001.

[8] Baddeley, A., "The episodic buffer: a new component of working memory?", Trends in cognitive sciences, 4: 417-423, 2000.

[9] Baddeley, A.D. and Hitch, G.J., "Working memory", In The Psychology of Learning and Motivation (Bower. G.A., ed.), pp. 47-89, Academic Press, 1974.

[10] Wise, R.J., Scott, S.K., Blank, S.C., Mummery, C.J., Murphy, K. and Warburton, E.A., "Separate neural subsystems within Wernicke's area", Brain, 124: 83-95, 2001.

[11] Hickok, G. and Poeppel, D., "Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language", Cognition, 92: 67-99, 2004.