# Speech shadowing as an elicitation technique in variation research: the case of the Italian mobile diphthongs

**Bart van der Veer*** and **Vincent J. van Heuven**
Universiteit Leiden Centre for Linguistics (ULCL), The Netherlands
*also at HIVT-HA, Antwerp, Belgium
E-mail: b.vanderveer@hivt.ha.be; v.j.j.p.van.heuven@let.leidenuniv.nl

## ABSTRACT

Italian features a stressed diphthong/unstressed monophthong alternation, which traditional grammars capture as the 'mobile diphthong rule': the rising diphthongs [jɛ] and [wɔ], historically related to Late-Latin mid-low stressed vowels, alternate with corresponding monophthongs due to some stress-shifting morphological operation. However, the written language features a growing generalization of these diphthongs in unstressed position. We aim to investigate to what extent this alternation (still) occurs in spoken Italian with use of the speech shadowing technique in combination with a phoneme restoration task. The experimental results show substantial variation, not only between and within speakers in the elicited production experiment but also between listeners in a perception task. We claim that variation in perception may account for the variation in speech production. From a methodological point of view, we conclude that speech shadowing eliciting spontaneous phoneme restoration is a reliable and useful technique in language variation research.

## 1 INTRODUCTION

The Italian 'mobile diphthongs' are the rising diphthongs [jɛ] and [wɔ], which are historically related to Late-Latin mid-low stressed vowels, that alternate with corresponding monophthongs due to some stress-shifting morphological operation, e.g. *muovo* ['mwɔvo] 'I move', *movimento* [movi'mento] 'movement' (Italian mid-low vowels are raised when unstressed). However, the written language features a growing generalization of these diphthongs in unstressed position, e.g. *muoviamo* [mwo'vjamo] 'we move'. In this paper, we aim to investigate whether spoken Italian features a similar tendency towards allomorphy reduction. A phoneme restoration task using the speech shadowing technique was carried out under laboratory conditions.

## 2 SPEECH SHADOWING

The shadowing task must allow us to record spontaneous vowel production while subjects remain unaware of the purpose of the experiment and, importantly, are not influenced by orthographic information (since we want to test spoken-language variation). In a speech shadowing task, subjects repeat ('shadow') speech (being delivered over headphones) as soon as they hear it, i.e. without waiting for the end of the stimulus utterance. By replacing the target vowel (e.g. [wo]/[o] in *c(u)ocevo* 'I was cooking') by noise, the technique allows us to combine on-line shadowing with a restoration task; see [1]. The phoneme restoration effect occurs when listeners (in this case shadowers) effortlessly and fluently 'fill in' the missing phoneme information in structures where the target is highly redundant, often without even being aware that the target is missing at all, cf. [2, 3]. The shadowing condition guarantees that restoration is performed under considerable temporal pressure, although shadowing latencies are generally less than a second, i.e. there is a time lag of less than one second between the moment the shadower hears a sound/syllable and the moment that he himself produces the same sound unit, see [4]. Since the fifties, shadowing tasks have been used in the domain of auditory word recognition by manipulating the original stimuli, see a.o. [1, 5, 6]. The application of the on-line shadowing technique combined with a phoneme restoration task is a new tool in the field of language variation research. The following sections describe the design, procedure and results of our experiment.

## 3 EXPERIMENT

### 3.1 SUBJECTS

Recordings were made with a group of ten speakers between the ages of 20 and 27, all university students, five males and five females, who were born and/or raised in the province of Pisa and who consider themselves speakers of Standard Italian and reported no speech defects. The informants were paid a fee.

### 3.2 MATERIALS

The selection of the target words was based on existing Italian words that theoretically are subject to a monophthong/diphthong alternation as a consequence of some stress-affecting morphological operation. In order to make a representative selection, the following variables in the stimulus materials were defined: (a) base lexeme (noun, adjective, verb); (b) type of diphthong (front/back); (c) morphological operation (derivation, inflection, diminutivization). A total of 16 target words (6 nouns, 6 adjectives and 4 verbs) were embedded in meaningful

 ISBN 1-876346-48-5 

carrier sentences, where they were preceded by their respective bases. These 18 sentences were preceded by five "triggering" sentences and further mixed with 20 filler sentences, with a structure similar to that of the target sentences but containing no potential diphthongs. An example is given in (1).

La pasta è buona, anzi è b(u)onissima. (1)
'The pasta is good, it is even very good.'

The 43 sentences were recorded through an AKG D.880 Emotion microphone and a 4-channel ultra-di pro preamplifier on a PC with the Maxi Studio Isis software. The PC was placed outside the recording booth. The reader was a female native speaker of Italian, aged 33, born and raised in the province of Florence. The 16 sentences containing the target words were recorded twice: once with and once without diphthongized syllable nuclei in the underlined words. Two additional recordings served as practice materials in the learning stage of the experiment. The test sentences were read at an average speaking rate of three words per second.

In a second phase, the original speech materials were downsampled and processed further with the Praat (version 4.0.1) speech processing software [7, 8] with 16 bit amplitude resolution and a 16 kHz sampling frequency. With the use of the information provided by the waveforms, every section containing a target nucleus was gated out and replaced by noise (Gaussian noise, high-pass filtered from 100 Hz with 50-Hz smoothing, created by Praat). An example of a waveform of a target word with noise is given in Figure 1.

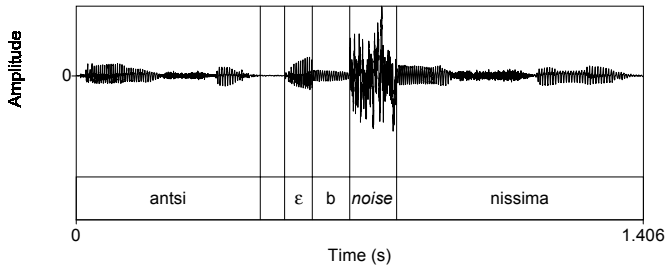


**Figure 1:** Waveform of the utterance *anzi è b(u)onissima* '(she) is even very good', where a section containing the target nucleus (underlined) has been replaced by noise.

The length of the noise burst was based on the mean length of a specific nucleus in a specific target word as produced by our reader – once as a monophthong and once as a diphthong. For example, in *buonissima* [wo] = 126 ms, in *bonissima* [o] = 98 ms, the mean duration = 112 ms; the noise was spliced into the recording such that it also replaced transitions from and to surrounding consonants. The resulting materials were then recorded from the PC onto a JVC minidisc (Crystal Gold) with a Sony Portable Minidisc Recorder MZ-R700.

### 3.3 PROCEDURE

The stimuli were delivered through headphones by the Portable Minidisc Recorder. The subjects were allowed to adjust the sound volume to a level they felt comfortable with, in order to perform the task. They were instructed to repeat the message as closely and fluently as possible, ignoring the noise bursts as well as they could. After the practice session, all speakers felt satisfied with their performance. The recordings were made in a sound-proofed booth at the Linguistic Laboratory of the Scuola Normale Superiore in Pisa. The utterances were recorded through a Sennheiser MD441-u microphone on a Casio DAT-DA2 recorder, which was placed outside the recording booth. The minidisc player was placed inside the sound-proofed booth and the subjects were asked to operate it when asked (press the 'start' and 'stop' buttons). This set-up was necessary, because it was not possible to deliver and record audio simultaneously from outside the booth; the portable minidisc player is silent enough for the purpose of this experiment. After the recording session, speakers were asked, if necessary, to repeat one or more sentences, if these had not been shadowed correctly in the first session.

### 3.4 PERCEPTUAL ANALYSIS

The corpus of 430 speech utterances (10 speakers × 43 sentences) was downsampled to 16 kHz and stored on hard disk with the Praat speech processing software. Only three relevant responses contained omissions and/or hesitations; these were excluded from further analysis. The (10 × 16) – 3 = 157 tokens were manually segmented and only labelled with an identification number. The transcription of the data was organized as follows. The target words (embedded in a small section of the original carrier sentence) were presented through headphones to five listeners: two Dutch phoneticians (the present authors) and three Italian phonetically naive native speakers. For each target word, printed on a score form in its two possible versions (either with monophthong or diphthong), the listeners had to make a binary choice: did the word contain a monophthong or a diphthong? They were allowed to listen to the tokens as often as they wanted. The task was repeated for each speaker.

## 4 RESULTS

The 157 stimulus items were scored as either a monophthong or a diphthong by five listeners, yielding a dataset of 1,085 responses.

### 4.1 AGREEMENT

Before analysing the effects of the experimental factors, we will first examine the agreement among the five listeners. For each stimulus any two listeners may come up with a value 0 (monophthong) or 1 (diphthong). The kappa coefficient is used to quantify the extent to which the scores of two listeners agree. Kappa is identical to a product-moment correlation coefficient for binary scores (0,1). When kappa equals 1, two listeners agree on the monophthong-diphthong nature of all stimuli, i.e., when listener A hears a monophthong, so does listener B, and when A hears a diphthong, B does as well. When kappa

equals 0 the decisions made by one listener have no relationship with those made by the other listener, i.e. the chances that listener B hears a monophthong or a diphthong are the same, irrespective of speaker A's decision.

| | BV (NL) | VH (NL) | CB (I) | VM (I) | IF (I) |
|---|---|---|---|---|---|
| **BV (NL)** | 1.000 | | | | |
| **VH (NL)** | .847 | 1.000 | | | |
| **CB (I)** | .324 | .207 | 1.000 | | |
| **VM (I)** | .404 | .303 | .542 | 1.000 | |
| **IF (I)** | .248 | .180 | .559 | .532 | 1.000 |

**Table 1:** Agreement among the 5 listeners, expressed in κ.

It is obvious from table 4.1 that the two (Dutch-speaking) phoneticians agree quite well in their decisions, with $\kappa = 0.847$ ($p < 0.001$). However, the three Italian native listeners have poorer kappa values, ranging between $\kappa = 0.532$ ($p < 0.001$) and $\kappa = 0.559$ ($p < 0.001$). Typically, the agreement between the Dutch and the Italian listeners is surprisingly low, with kappa's between $\kappa = 0.180$ ($p < 0.001$) and $\kappa = 0.404$ ($p < 0.001$). Given the rather poor correlation coefficients between the Dutch and the Italian listeners, we decided to examine the perceptual judgements in more detail.

Figure 2 presents the mean diphthong scores for the five listeners. The two Dutch listeners have diphthong scores of 0.47 and 0.52, whilst the three Italians' scores range between 0.76 and 0.82. The effect of listener is highly significant by a oneway Analysis of Variance, $F(4, 780) = 20.4$ ($p < 0.001$). Scheffé post hoc analyses for contrasts ($\alpha = 0.05$) show that the scores for the Dutch listeners do not differ from each other but do differ from those of each of the Italians, which do not differ amongst each other.

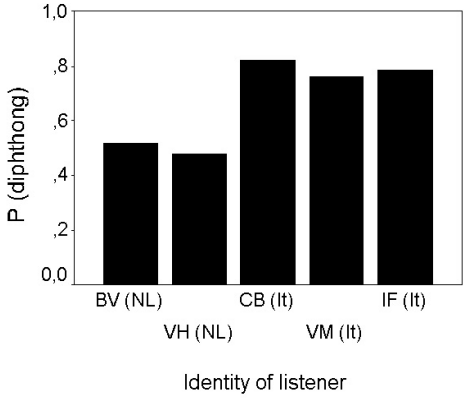


**Figure 2:** Number of diphthongs perceived (proportion) broken down by listener.

Apparently, the Italian native listeners are much more prone to hear a diphthong than the Dutch phoneticians. In roughly two-thirds of the cases where the Dutch listeners report a monophthong, the Italians perceive a diphthong (71% for VH and 65% for BV). It is unclear at this stage whether the difference between the Dutch and the Italian listeners is a matter of response bias (possibly induced by orthographic practice), or whether the Italians attend to subtle diphthongization cues that elude the Dutch phoneticians.

### 4.2 EFFECTS OF EXPERIMENTAL FACTORS

Figure 3 presents the probability (proportion) of perceiving a diphthong across all five listeners, broken down by morphological operation (inflected forms, derivations and diminutives) for /(j)e/ and /(w)o/ forms.

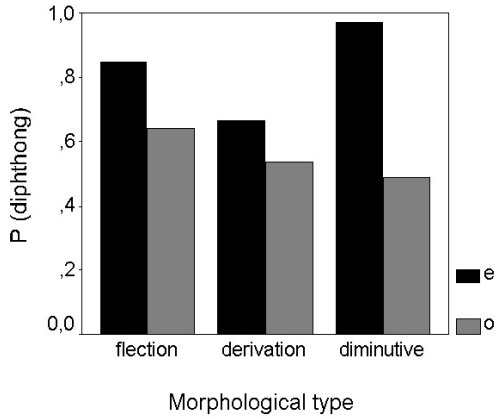


**Figure 3:** Proportion of perceived diphthongs /je/ and /wo/ broken down by three morphological operations.

The incidence of diphthongs is higher for front /e/ (79%) than for back /o/ (55%); this effect is large and highly significant, $F(1, 9.003) = 49.5$ ($p < 0.001$). More diphthongs are reported in inflected words (75%) than in derivations (60%), with somewhat higher diphthong scores for /e/ than for /o/. However, in diminutives the probability of perceiving a diphthong drops further (49%) when the vowel is /o/ but increases to a dramatic 97% for front vowels /e/. As a result of this the effect of morphological operation is significant, $F(2, 18.014) = 5.6$ ($p = 0.013$) as is the interaction between vowel type and morphology, $F(2, 18.020) = 8.4$ ($p = 0.003$).

## 5 CONCLUSION AND DISCUSSION

The primary aim of this chapter was to investigate to what extent the diphthong - monophthong alternation, in Italian grammars captured as 'the mobile diphthong rule', still occurs in spoken Standard Italian. Clearly, the allomorphy reduction, attested in the written language, also takes place in the spoken language: no alternation occurred in 70% of the tested word pairs (on average). However, there is substantial variation in the experimental results. The elicited production based on the shadowing technique proved extremely useful in demonstrating this variation between and within speakers. Variation also occurred between listeners in the perception task. In this section, we will provide an explanation for the variation that was found in the listening task and subsequently claim that variation in perception may account for the variation in speech production.

In our discussion of the perceptional variation, we will be concerned with the acoustic description of the target vowels,


ISBN 1-876346-48-5 

based on formant frequencies. It is commonly known that the frequencies of the first and the second vowel formant ($F_1$ and $F_2$) correspond with a traditional vowel quadrilateral, such as the IPA vowel chart (cf. [9]). A prototypical [i] (and its semivowel allophone [j]) has higher $F_2$ values and lower $F_1$ values than front mid-vowels [ɛ] and [e]; $F_1$ and $F_2$ values of the back rounded vowel [u] (and the corresponding semivowel [w]) are both lower than those of the back mid-vowels [ɔ] and [o]. Hence, the first and second formant tracks may serve as a cue for the distinction between a monophthong and a diphthong.

In front rising diphthongs, $F_1$ starts low and $F_2$ high ([j]) and the distance will decrease towards the temporal centre of the front mid-vowel, whereas in back rising diphthongs both $F_1$ and $F_2$ start low and both rise in a parallel movement, see figure 4.

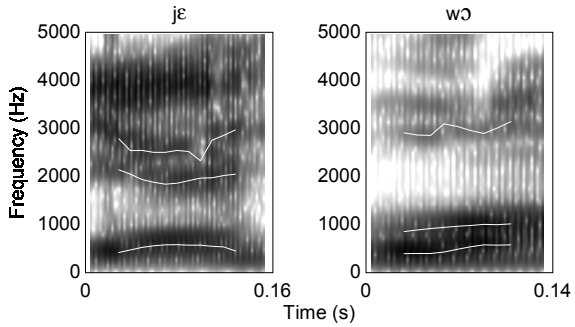


**Figure 4:** Spectrograms of stressed [jɛ] and [wɔ], pronounced by a male and female speaker, respectively. Formant tracks for $F_1$, $F_2$ and $F_3$ are indicated in white.

As can be concluded from figure 4, back rounded diphthongs have $F_1$ and $F_2$, which are typically close together in frequency. In their front counterparts, $F_1$ and $F_2$ are more separated and make anti-parallel movements. Classic experiments have shown that formants that are close together in frequency may combine into a single perceived peak (see [9] and references therein). This fact may suggest that gliding along the close-open dimension in rising back diphthongs, especially when unstressed and therefore relatively short, is more difficult to perceive than gliding in rising front diphthongs, because in the former the low- frequency formants may fuse into a single peak and no transition from one vowel to another is perceived. It can therefore be assumed that $F_1$ and $F_2$ frequency values constitute subtle diphthongization cues, with back diphthongization being more subtle than front diphthongization, to which listeners (and particularly groups of listeners with different native languages) attend differently.

Importantly, if front diphthongs are assumed to be perceived more clearly than back diphthongs, we may hypothesize that the two groups of listeners agree better in their perceptions of front diphthongs than in that of back diphthongs. To test this hypothesis, the collected data were submitted to an ANOVA with listener nationality and vowel type as fixed factors, and with speaker as a random factor. This analysis indicates highly significant effects for vowel type, $F(1, 9.007) = 36.6$ ($p < 0.001$), and for nationality $F(1, 9.010) = 57.8$ ($p < 0.001$). Also highly significant is the interaction between nationality and vowel type, $F(1, 9.012) = 36.0$ ($p < 0.001$). The probability that a rising back diphthong is perceived is .25 for the Dutch listeners against .75 for the Italians; for rising front diphthongs these values are .74 for the Dutch and .82 for the Italian listeners. These results confirm our assumption that the Dutch subjects are much less sensitive than the Italians in their perception of diphthongization in back vowels than in front vowels.

If we assume that the 'mobile diphthong'/monophthong alternation in Italian is reduced in favour of the diphthong in both stressed and unstressed position (suggested by the written language and confirmed by our experiment), we still need to account for the fact that this reduction process is unbalanced, rising front diphthongs being far more generalized in unstressed position than rising back diphthongs. Since we now know that front diphthongization is more easily perceived than back diphthongization, it is not surprising that the tendency towards preservation of the diphthong in all stress positions is stronger for front diphthongs than for back diphthongs. In general, we would claim that the phonological process of allomorphy reduction (or paradigm uniformity) is strongly determined by phonetic nuances.

## REFERENCES

[1] Heuven, V.J. van, "Effects of stress and accent on the human recognition of word fragments in spoken context: gating and shadowing," in *Proceedings of the 7th FASE/Speech-88 Symposium*, W.A. Ainsworth and J.N. Holmes, Eds. pp. 811-818. Edinburgh: Institute of Acoustics, 1988.

[2] Warren, R.M., "Perceptual restoration of missing speech sounds," *Science* 167, pp. 392-393, 1970.

[3] Samuel, A., "Phoneme restoration," *Language and Cognitive Processes* 11, pp. 647-654, 1996.

[4] Marslen-Wilson, W.D., "Linguistic structure and speech shadowing at very short latencies," *Nature* 244, pp. 522-523, 1973.

[5] Cherry, C., "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America* 25, pp. 975-979, 1953.

[6] Marslen-Wilson, W.D., "Sentence perception as an interactive parallel process," *Science* 189, pp. 226-228, 1975.

[7] Boersma, P. and D. Weenink, Praat 4.0.1 a system for doing phonetics by computer, 1996. [www.praat.org]

[8] Boersma, P. and V.J. van Heuven, "Speak and unSpeak with Praat," *Glot International* 5, pp. 341-347, 2001.

[9] Hayward, K., *Experimental phonetics*. Harlow: Longman, 2000.