Perception of gesturally distinct consonants in Persian

Reza Falahati
Laboratorio di Linguistica
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italy
reza.falahati@sns.it

Chiara Bertini
Laboratorio di Linguistica
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italy
chiara.bertini@sns.it

Abstract

This study explores the sensitivity of the individuals to the residual gestures remaining after the simplification of consonant clusters. Three sets of target stimuli having full, reduced, and zero alveolar gestures along with the control stimuli were used in a perceptual identification task. The results of the experiment showed that subjects reliably distinguished the three target sets with varying residual gestures from the control. The results also showed that the degree of residual gestures affects the rate of [t] perception by the subjects; however, this was not statistically significant. The results are discussed in the context of different theories of speech perception.

1 Introduction

This study investigates the perception of three categories of consonant clusters that are perceptually similar but gesturally distinct. In Persian, word-final coronal stops are optionally deleted, when they are preceded by obstruents or the homorganic nasal /n/. For example, the final clusters in the words /ræft/ "went", /duχt/ "sew" and /qæsd/ "intension" are optionally simplified[1] in fast/casual speech, resulting in: [ræf], [duχ], and [qæs], respectively. The articulatory study conducted on this process in Persian by Falahati (2013) has shown that the gestures of the deleted segments are often still present. More specifically, the findings showed that of the clusters that sounded simplified, some had no alveolar gesture, some had gestural overlap that masked at least some of the acoustic information for [t], and some had reduced alveolar gestures. The current study tests listeners' sensitivity to these three types of /t/ realizations.

[1] The term "simplification" is used here for the acoustic and perceptual consequence of apparent coronal consonant deletion, regardless of whether there is a residual articulatory gesture.

2 Background

Choosing the basic units or building blocks by which the phenomena in a discipline could be explained is fundamentally important. Due to the "complex" nature of language, there is no consensus among linguists as to the nature of this basic unit in the field. The controversy over choosing the building blocks extends to the domain of speech perception where different models have postulated various basic units of processing and storage.

In general, there are two major theoretical approaches to speech perception: gesturalist theories versus auditory and exemplar theories. The two main gestural theories of speech perception are Motor Theory and Direct Realism (MT and DR, henceforth). In motor theories, the intended phonetic gestures of the speaker are considered to be the objects of speech perception. These gestures are "represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations" (Liberman and Mattingly 1985, p. 2). The main motivation for choosing such basic unit by MT, among other factors, is mainly because of patterns where different acoustic cues could give rise to the same phonetic percept or where variant phonetic percepts were found for the same synthetic speech across different contexts (Delattre et al., 1955, 1964; Liberman 1957; Liberman and Mattingly 1985). Despite of the fact that this theory has gone through significant changes from its inception, all the versions share the idea that the objects of speech perception are articulatory events rather than acoustic or auditory events.

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

An intended gesture is produced by a number of muscles that act in concert sometimes ranging over more than one articulator. For instance, constriction needed for producing coronal stops involves the action of the tip/blade of the tongue and the jaw; however, such a constriction is considered one gesture. According to MT, the orchestration among gestures is quite systematic and listeners can use the systematically varying acoustic cues for coronal stops as information to detect the related consonant gestures.

MT assumes a biological link between perception and production. According to this perspective both speech perception and speech production share the same set of invariants and are governed by auditory principles. "The motivation for articulatory and coarticulatory maneuvers is to produce just those acoustic patterns that fit the language-independent characteristics of the auditory system" (Liberman and Mattingly, 1985, p. 6). The acoustic signal only serves as a source of information about the gestures. It is the gestures which define the phonetic category.

The other main gestural theory to speech perception is direct realism. Both DR and MT share the claim that listeners to speech perceive vocal tract gestures. However, in DR it is the phonological gestures of the vocal tract, rather than the intended gestures, which are the perceptual objects (Fowler 1981, 1984, 1996). According to DR, "the temporal overlap of vowels and consonants does not result in a physical merging or assimilation of gestures; instead, the vowel and consonant gestures are coproduced. That is, they remain, to a considerable extent, separate and independent events..." (Diehl et al., 2004, p. 153). If we could extend this to the gestures of two adjacent consonants, one should expect that the gestures related to them also remain separate and distinct from each other.

In contrast to gestural theories, the auditory theories assume that speech sounds are perceived via general cognitive and learning mechanisms. In this view, speech is not special and listeners do not perceive gestures. The auditory approach to perception mainly considers general auditory mechanisms responsible for perceptual performance. According to this view, the speech and nonspeech stimuli do not invoke a special or speech-specific module. Gestures have no mediatory role as to the perception of speech sounds in this approach. Listeners use multiple imperfect acoustic cues in order to categorize the complex stimuli with structured variance (Diehl et al., 2004). According to this approach, the phonological representations are assumed to be speaker independent and they are associated with each word in the listener's mental lexicon. The proponents of this approach take, for example, categorical perception of non-speech sounds or categorical-like perception by non-human animals as evidence for their argument. They also consider some of the cross-linguistic sound patterns and the "maximal auditory dispersion" in vowel systems as further support for their claim (Ohala 1990, 1995).

Exemplar theories form another approach to speech perception where words and frequently-used grammatical constructions are represented in memory as large sets of exemplars containing fine phonetic information. Listeners are sensitive to phonetic details existing in the speech signal. In such a speech perception model, a mechanism is needed for gradiently changing the lexical representations over time. In order to do so, the perceptual system must be capable of making fine phonetic distinctions (Johnson 1997).

These different approaches to speech perception have been tested in different studies. Beddor et al., (2013), for example, used eye-tracking to assess listeners' use of coarticulatory vowel nasalization as that information unfolded in real time. In the experiment, subjects heard the nasalized vowels with two different time latencies. The prediction was that subjects will fixate on the related image sooner when they hear the nasalized vowel earlier. The results showed that listeners use relevant acoustic cues, which was argued to allow listeners to track the gestural information. Nalon (1992) in an identification task tested whether participants could identify different degrees of velar assimilation. He used four different articulation types called full alveolar, residual alveolar, zero alveolar (i.e., full assimilation to the following velar), and nonalveolar (i.e., velar in underlying representation). The results of his study showed that the participants perceived full alveolar tokens with 100% accuracy with /d/ responses while less than half the tokens with residual alveolar were identified with /d/ responses. In another study, Pisoni showed that the nonspeech analogs of VOT stimuli are perceived categorically. Similar studies like this were taken as evidence against MT which claimed categorical perception as a specific feature of the speech mode of perception.

In this study, I will use three sets of simplified consonant clusters which are auditorily similar but gesturally different. The consonant clusters (i.e., C_1C_2#) happen in the coda of the words followed by another word which also starts with a consonant, therefore giving us three consonants in a row in an intervocalic environment (i.e., $V_1C_1C_2$#C_3V_2). The prediction is that if subjects are sensitive, they should have different judgment for the stimuli. The stimuli set with no coronal gesture is expected to show the same pattern as the control (with zero coronal gesture in the underlying representation). The stimuli with overlapped gestures and reduced gestures are predicted to show a pattern different both from control and the stimuli with zero residual gestures. The following section introduces the methodology of the study.

3 Methodology

3.1 Participants

Thirty-two Persian-speaking students from the Università di Pisa and Sant'Anna, seventeen females fifteen males, aged 18-38 participated in this study. The results of eight of them are not considered for analysis because they reported to be bilinguals and mainly used a language rather than Persian at home or with their close friends. This resulted in twenty-four, twelve females twelve males. None of them reported any hearing problem.

3.2 Stimuli

Three sets of target words varying in only the degree/amount of alveolar residual gestures and one control stimuli set were used in the experiment. The three target categories are mainly the same except for the degree of alveolar residual gestures. Target Full_G category has full coronal gesture but has overlap hence marked with two superscript [^{tt}]. Target Partial_G category has partial residual gesture marked via superscript [^{t}] whereas Target Zero_G has no gestural leftover. The stimuli in the control are used as the baseline since they don't have any underlying coronal stop in the coda position of the first word. Some examples of the target and control words are given below:

Target Full_G: [æχ^{tt} kɑ], [æf^{tt} bæ], [uf^{tt} bɑ]
Target Partial_G: [æχ^{t} kɑ], [æf^{t} bæ], [uf^{t} bɑ]
Target Zero_G: [æχ kɑ], [æf bæ], [uf bɑ]
Control: [æχ ke], [æf bæ], [uf bɑ]

The four sets of target and control nonwords presented above are the excised tokens taken from the full words presented below:

Target: /sæχt kɑr/ "hard-working", /næft bærɑje/ "oil for", /kuft bɑʃeh/ "be cheap"

Control: /næχ ke/ "thread that", /sæf bærɑje/, "cue for" / mæruf bɑʃeh / "be famous"

3.3 Procedure

All the participants listened to forty stimuli (10 stimuli in each category) with eight repetitions. (total of 320 tokens) in a sound booth located at the linguistics laboratory in Scuola Normale Superiore. The software Presentation was used to present the stimuli to the listeners as an identification task. The participants were asked to listen very carefully and decide as quickly as possible whether it is likely that there has been a [t] at the end of the first part of each stimuli. For each stimulus, the participants were asked to press either the green or the blue button on a Cedrus response pad. On the screen of a computer, listeners could also see "T" or "NO T" corresponding to the response buttons. The stimuli were shuffled and presented in blocks in a way that participants could either begin by hearing all the tokens with [f] or [χ]. They also had the choice of taking a break after listening to every 80 tokens. All the participants received a short training before the start of the experiment. The following section contains the results of the study.

4 Results

The main goal of this study is to test listeners' sensitivity to different degrees of residual gestures remaining after the simplification of consonant clusters. The response type and reaction time are the dependent variables in this study; however, only the results related to response type are presented here. Figure 1 below shows the perception rate of [t] by all subjects

across the four conditions. According to this, the subjects show the highest rate of [t] perception in tokens with full alveolar gesture (i.e., 59.69%) and the lowest for the ones in the control (i.e., 36.09%). The condition with partial alveolar gestures shows the rate of 56.20% which is very close to the full condition. The stimuli in zero alveolar condition show an intermediate level between the control and the other two target conditions with the rate of 49.84%. This shows almost a similar pattern between the two target conditions with full and partial gestures, an intermediate situation for the target condition with zero gesture, and a pattern for the control which is different from the three target conditions.

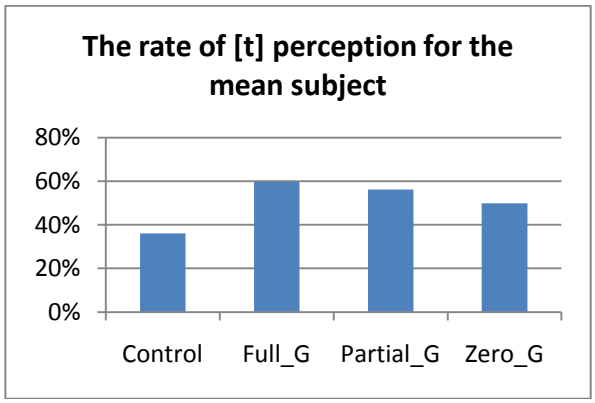

Figure 1: The Rate of [t] Perception by all Subjects

In order to examine the relation between the two categorical variables in the study, namely the response type and stimuli condition, a Pearson chi-square test was run. The null hypothesis is that there is no relation in the [t] perception and the four conditions in the study. The results of the test with [t] perception as the dependent variable found significant main effect of conditions χ^2 (3, N = 960) = 46.2, $p < 0.001$. This shows that there is a significant relation between the stimuli conditions and response type. In order to determine whether the difference in the perception of [t] across four categories is really significant or it is due to chance variation, a column proportions test was performed. This test uses z-test to make the comparisons. The result showed that the perception of [t] in the control was significantly different from the all target categories. The next section presents the discussion and concluding remarks of the study.

5 Discussion and Conclusion

This research investigated listeners' sensitivity to three types of /t/ realizations as target and compared the results with the control. The target categories included simplified consonant clusters with full, partial, and zero alveolar gestures. The stimuli used as the baseline in the control had no alveolar gesture in the underlying form. The general results of the study showed that subjects reliably distinguished the three target sets with varying residual gestures from the control. This could be due to more similarity in tongue configuration in realizing these varying degrees of coronal stop articulation compared to the control condition where there is no alveolar gesture in the underlying form. Any articulatory modification is expected to trigger acoustic changes. The acoustic results of the stimuli used in this study by Falahati (2013) showed no significant difference between the simplified tokens (i.e., the three target sets with varying degrees of residual gestures labeled all together as simplified) and control tokens. The acoustic parameters used in the analysis were V_1 duration, consonant clusters duration, and formant transitions. Despite of the fact that the results did not show any significant difference between simplified and control conditions, the duration of V_1 and consonant clusters in the simplified condition was always higher than the control condition. It could be the case that these acoustic cues, although not very strong, are enough for human's auditory system to trigger the presence of a segment.

The results of the current study also showed that participants perceived almost 36% of the tokens with no underlying coronal stop as having [t]. This is very similar to the results of the study reported by Nalon (1992) where 20% of the control nonalveolar tokens were perceived as having [d]. In his study, however, the control tokens showed similar pattern to that of the target with zero alveolar (i.e., full assimilation). He attributes this to both subjects' natural language experience as well as the inherent ambiguity in the stimuli. He states that subjects are "willing to "undo" its effects" and therefore, in the case of the current study, report coronal stops even where there is no evidence for them.

The results of our study also showed that participants perceived more [t] in the tokens with full and partial alveolar gestures compared to the ones with zero alveolar gestures. The difference

between the three categories, however, did not reach the significance level. Such result could shed more light on the theories of speech perception discussed earlier in this paper. In order to discuss this issue, first we need to further explore the nature of the three categories in the target stimuli. From the three groups in the target stimuli, one group categorically had no alveolar gesture while the other two had different degrees of the gesture either as a result of overlap or reduction. We argue that the gradient gestural reduction and overlap are due to low-level phonetic and mechanical reasons while the categorical deletion, which results in tokens with zero gestures, is caused by the cognitive system. In the former groups, speakers neither intend to reduce nor plan to overlap gestures while the latter process is intended by the speaker.

According to MT and DR, listeners' target in speech perception is the intended or phonological gestures. Therefore, the overlapped and reduced stimuli should show different perceptual pattern compared to the stimuli with no residual gesture. The results in this study did not show a striking difference between these three target sets. The existence of acoustic cues pertaining to the presence of gestures is a prerequisite to their perception by the listener. If distinguishing acoustic details could be found between these three categories, then this would not support the gesturalist approach to speech perception. However, with the current results, such a claim cannot be made. Further acoustic analysis between these three target sets is needed to examine this idea further.

The findings in our experiment could be best explained by referring to exemplar models of speech perception. In such models, the lexical representations of words change in a gradient way over time. This is due to the nature of some phonological processes in languages which are not categorical. According to this view, the perceptual mechanism is capable to make fine phonetic distinctions. However, it is the mapping between the gradient stimuli and the auditory system which fails and does not result in nonvariant forms.

The lack of such a one-to-one mapping will bring variation across subjects in the speech community. The degree of such variation is determined by the *amount* of individual's exposure to the *specific* variants. A closer look at the results for individual subjects showed that all twenty-four participants in the study could fall into three or four dominant patterns based on their perception of [t]. The variation across individuals regarding speech perception could be a good source of information for the specialists in the field. Moreover, the degree to which an individual's speech production could map to his/her perception is an interesting topic which remains to be explored.

Acknowledgments

We are very grateful to Patrice Beddor for her comments and suggestions on this study.

Reference

Patrice S. Beddor, Kevin B. McGowan, Julie Boland, Andries W. Coetzee, and Anna Brasher. 2013. The perceptual time course of coarticulation. *Journal of the Acoustical Society of America*, 133, 2350-2366.

Pierre Delattre, Alvin M. Liberman, and Franklin S. Cooper. 1955. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27, 769-773.

Pierre Delattre, Alvin M. Liberman, and Franklin S. Cooper. 1964. Formant transition and loci as acoustic correlates of place of articulation in American fricatives. *Stud. Linguist*. 18, 104-121.

Randy L. Diehl, Andrew J. Lotto, and Lorri L. Holt . 2004. Speech perception. *Annual Review of psychology*. 55, 149-179.

Alvin M. Liberman and Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. Cognition, 21: 1-36.

Reza Falahati. 2013. *Gradient and Categorical Consonant Cluster Simplification in Persian: An Ultrasound and Acoustic Study*, Ph. D Dissertation, University of Ottawa, Ottawa.

Carol C. Fowler. 1981. Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, 46, 127-139.

Carol C. Fowler. 1984. Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, 36, 359-368.

Carol C. Fowler. 1996. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99, 1730-1741.

Keith Johnson. 1997. Speech perception without speaker normalization: an exemplar model. In K. Johnson and J. W. Mullennix (eds.), *Talker variability in speech processing*, 145-165. San Diego: Academic Press.

Alvin M. Liberman and Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition*, 21: 1-36.

Francis Nalon, 1992. The descriptive role of segments: evidence from assimilation. In G. J. Docherty and R. Ladd (eds.), *Papers in Laboratory Phonology IV*, 261-289. Cambridge: Cambridge University Press.

John Ohala. 1990. Respiratory activity in speech. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling*, 23-53. Netherlands: Kluwer Academic Publishers.

John Ohala. 1995. The perceptual basis of some sound patterns. In B. Connell and A. Arvaniti (eds.), *Phonology and phonetic evidence, Papers in Laboratory Phonology IV*, 87-92. Cambridge: Cambridge University Press.