Manuscript submitted to Cerebral Corrtex

Listener-speaker perceived distance predicts the degree of motor contribution to speech perception

Authors:

Eleonora Bartoli[1,§], Alessandro D'Ausilio[1,§], Jeffrey Berry[1,2], Leonardo Badino[1], Thomas Bever[2,#], Luciano Fadiga[1,3,#]

Affiliations:

[1] IIT - Italian Institute of Technology. RBCS – Robotics, Brain and Cognitive Sciences department

Via Morego, 30, 16163 – Genova, Italy

[2] University of Arizona Cognitive Science Program

1503 E. University Blvd., 85721-0068 - Tucson, AZ, USA

[3] University of Ferrara, DSBTA – Section of Human Physiology

Via Fossato di Mortara, 17/19, 44100 – Ferrara, Italy

[§] These authors equally contributed to the work

[#] These authors equally contributed to the work

Corresponding Author: Alessandro D'Ausilio

IIT - Italian Institute of Technology, RBCS – Robotics, Brain and Cognitive Sciences department.

Via Morego, 30 - 16163 - Genova, Italy

t: + 39 010 71781975; f: + 39 010 7170817

e-mail: alessandro.dausilio@iit.it

Running Title: Motor perception of speech: inter-speaker distance

Abstract

Listening speech sounds activates motor and premotor areas in addition to temporal and parietal brain regions. These activations are somatotopically localized according to the effectors recruited in the production of particular phonemes. Previous work demonstrated that transcranial magnetic stimulation (TMS) of speech motor centers somatotopically altered speech perception, suggesting a role for the motor system. However, these effects seemed to occur only under adverse listening conditions, suggesting that degraded speech may stimulate listeners to adopt unnatural neural strategies relying on motor centers. Here we investigated whether naturally occurring inter-speaker variability, which did not affect task difficulty, made a speech discrimination task sensitive to TMS interference. In this paradigm, TMS over tongue and lips motor representations somatotopically altered the discrimination time of speech. Furthermore, the TMS-induced effect correlated with listeners' similarity judgments between listeners' and speakers' speech productions. Thus, the degree of motor recruitment depends on the perceived distance between listener and speaker. This result supports the claim that discriminating others' speech pattern requires the contribution of the listener's own motor repertoire. We conclude that motor recruitment in speech perception can be a natural product of discriminating speech in a normally variable and unpredictable environment, not merely related to task difficulty.

Keywords

inter-speaker variability, motor system, motor theory, speech perception, TMS

Several influential theories of speech discrimination have assumed that a listener's representation of a speaker's articulatory gestures participates in speech perception (Liberman and Cooper 1967; Stevens and Halle, 1967; Fowler 1986; Callan et al. 2010). In contrast, several competing theories have suggested that a purely sensory analysis is sufficient for classification (Diehl et al. 2004).

Sensorimotor theories predict that speech classification is achieved in part through the listener's internal recapitulation of the phono-articulatory gestures produced by the speaker. In support of that, motor and premotor activations have been reported while listening to speech sounds (e.g. Fadiga et al. 2002; Watkins et al. 2003; Pulvermüller et al. 2006; Callan et al. 2004; Binder et al. 2004; Wilson et al. 2004; Shahin et al. 2009; Londei et al. 2010).

Transcranial Magnetic Stimulation (TMS) has been used to alter activity in different motor centers to demonstrate their localized causal involvement in speech discrimination tasks (Meister et al. 2007; Sato et al. 2009; Möttönen and Watkins, 2009; D'Ausilio et al. 2011a). However, most of the prior TMS studies used artificially degraded speech. When stimuli were presented without noise the effects found in noisy/degraded listening conditions were not replicated. (D'Ausilio et al. 2012a).

This observation has led some researcher to argue that the role for the motor centers is only modulatory and appears only in extreme circumstances (Hickok et al. 2011). A more direct criticism is that the stimulation of motor centers affects speech perception because of the use of "unnatural stimuli, eliciting atypical strategies based on atypical motor activities" (see Gernsbacher in Gallese et al. 2011).

In response to these criticisms it should be noted that, in nature, speech stimuli are almost always noisy and degraded, since in normal conditions many factors induce a significant variation in speech stimuli. These factors can be external to the speech signal, as noisy environments, for example, driving in traffic, travelling in an airplane, sitting in a tapas bar or walking next to the seashore. Features that stress the perceptual system can also be intrinsic to the signal itself, because of different acoustic/phonetic patterns unique to each speaker. For example, the gender of the speaker, age and other variations in the vocal tract shape are pervasive sources of these inter-individual variations, along with speed and regional accent of pronunciation (Adank et al. 2009; Dupoux and Green, 1997; Floccia et al. 2006). Secondly, even if motor effects were experimentally observable only with degraded stimuli, it would not prove that the motor system plays a minor or exceptional role. Thus, even the existing studies leave open the possibility that motor activation occurs in all cases of speech discrimination, but difficult judgments are necessary to make it experimentally observable.

In the present experiment we introduced a common natural source of speaker variation, inter-individual differences. Inter-individual speech variability naturally increases the number of features not strictly relevant to the discrimination task, thus representing a possible confound without special degrading of speech. In this study we applied online TMS stimulation to either lip or tongue motor representation while subjects discriminated syllables, with initial bilabial or dental consonants, spoken by 15 different speakers (following the method in D'Ausilio et al. 2009).

Our overarching goal was to develop evidence that an individual uses his/her own production system output as part of speech discrimination, matching it to the speaker's input. If a perceiver uses an internal template based on his own production

system, we expect to detect that the modulation on the motor system recruitment during perception is a function of the goodness of fit of the template of the perceiver with respect to the perceived speech features. Thus, similarities between the perceivers' template and the speakers' features should play a role in the degree of motor system recruitment during speech perception. To pursue this prediction, we measured subjective (perceptual judgments) and objective (acoustic measurement) distances between the experimental stimuli and the TMS subject's own vocal productions. The goal was to verify whether inter-speaker variability was a factor in the causal recruitment of the motor system and if the TMS-induced effects depended on subjective or purely acoustic distances between the stimuli and the subjects' own vocal production.

Materials and Methods

Subjects

Twelve right-handed healthy subjects (6 Males and 6 Females, mean age and standard deviation: 27.8 ± 5.6 years;) took part in the experiment. Subjects first participated in the TMS experiment and then in the multidimensional scaling (MDS) study. The two sessions were performed on different days. Participants were paid for their participation; the local ethics committee approved all procedures.

Stimuli database

Audio recordings of 30 Italian native speakers (15 Males; mean age: 29.1; ± 3.1 years) were collected. The speakers read the syllables presented one by one in randomized order on a laptop screen using a PsychToolbox script running in Matlab (Mathworks, Inc.). Stimuli consisted of eight different CV syllables derived from the

combination of four consonants ([b], [p], [t], [d]) and two following vowels ([a], [i]). Consonants were chosen with 2 different points of articulation (labials, dentals) and different voicing (voiced, voiceless) to match the experimental design of D'Ausilio et al. (2009). The two different vowels were included to give more variability to the stimuli.

Each syllable was repeated three times. The stimuli were recorded with a professional microphone with the freeware program Audacity (http://audacity.sourceforge.net/). The recordings were processed to reduce background noise using Audacity's built-in noise reduction tool (by a combination of noise reduction and frequency smoothing), down-sampled to 16 kHz, and segmented into 550 ms segments. Stimulus onset was aligned at 150 ms for every segment. Sound levels were then normalized using Vaill's normalization tool (available at http://normalize.nongnu.org) and a second audio channel was parallel to each stimulus, with a sound spike 100ms before stimulus onset. This channel was used to trigger TMS stimulation, and was not heard by the subject.

Stimuli selection

In order to find the subset of stimuli with maximum differences between speakers, we computed pairwise Euclidean distances between principal components of mel-frequency cepstral coefficients (MFCC, widely used in speech research to analyze acoustic variability), and pairwise Euclidean distances between the formants f1 and f2 of the stimuli. For MFCCs, each segment was processed from 100 ms before the onset of the vowel to 50 ms after the onset of the vowel. This captured the voicing quality of the consonant together with its release and formant transitions. MFCCs were calculated using 16 ms windows with 5 ms overlap, resulting in 14 windows with 13

MFCCs each (a 182 dimensional vector for each segment). Calculating Euclidean distances between such large vectors leads to similar distances between every pair of tokens, limiting their usefulness (Beyer et al. 1999). To overcome this problem, we reduced the dimensionality of the MFCC vectors using standard principal components analysis (PCA), projected the data onto the first 15 principal components, and measured the Euclidean distances in the reduced space. The first 15 components accounted for 94.78% of the total variation of the MFCCs, which allowed for a reasonable approximation while avoiding the problems of high dimensionality. All calculations for MFCC, PCA, and Euclidean distances were performed using the SciKits library for Python (http://scikits.appspot.com/).

The second measurement was related to formants, because they are strictly linked to dynamic configurations of the articulators. We used the first and second formant values measured at the midpoint of the vowel. Formant values were calculated automatically using the Forest Tool which is part of the Emu Speech Database System (http://emu.sourceforge.net/). Formant distances and MFCC-PCA distances were each scaled to values between 0 and 1. For each stimulus, the formant and MFCC-PCA distances were summed to produce a final distance score. We then selected a subset of stimuli that maximized the combined distance scores of the subset. A set of 6 segments from different speakers was selected for each syllable type (including 3 female and 3 male speakers), by maximizing the distance scores between the 6 tokens (see Fig 1). Maximization was done first by randomly choosing a segment as the point against which the distance scores would be measured. Segments were selected in order to maximize the sum of the distances from the previously selected tokens. This was repeated until the 3 tokens for each gender were selected, with only one token for each speaker. The selected set was saved together with the total distance score for all

the selected segments. Thus, the final set of 6 segments was the set with the highest total distance score (Figure 1). The entire procedure was repeated for each of the 8 syllables. The final set of stimuli consisted of 48 tokens (6 segments x 8 syllables) Since we did not constrain the selection of the speakers but only the maximization of the distance score, the final set of stimuli obtained comprised a subset of 15 (8 males) out of the initial 30 speakers sample. We then ran a Common Principal Components Analysis test (Flury 1988) to compare the covariance matrices of the two classes of stimuli. Using the "Jump-up" method described by Phillips & Arnold (1999) against unrelated/arbitrary structure we found that the 2 covariance matrices were proportional to each other (Chi-Square, 253.02; DoF, 153; $p < 0.00001$), but not equal (Chi-Square=295.00; DoF, 152; $p < 0.00001$), thus showing that the difference in covariance was statistically significant. This analysis was performed using Phillips' CPC software (http://pages.uoregon.edu/pphil/programs/cpc/cpc.htm). Finally, we computed the Euclidean L1-norms for each covariance matrix (the maximum absolute column sum), showing that tongue-produced phonemes had larger variance than lips-produced ones (4.757 and 4.619). This fact comes to no surprise since the range of labial choices is much simpler than the range of tongue possible configurations from a biomechanical point of view (Beautemps et al. 2001), which in turn is reflected in a greater number of distinctive places of articulation when considering the tongue with respect to the lips (Chomsky and Halle 1968; Clements 1985).

INSERT FIGURE 1 ABOUT HERE

General Procedures

Participants were asked to listen to each audio trace presented through earphones and to identify the consonant in the CV syllable as fast as possible. They responded with their left hand (ipsilateral to the stimulated hemisphere to avoid interference between hand response programming and TMS stimulation) by pressing on a 4-buttons response box the button corresponding to the consonant. Reaction times and accuracy were recorded using E-prime® software (Psychology Software Tools, Inc.). The experimental session was divided in two blocks, one for each site of TMS administration (lip or tongue primary motor cortex). For each of the two blocks, the 48 audio files were repeated twice in randomized order, with and without TMS, resulting in 96 trials for each block, with a total of 192 trials for each subject. The study lasted about 60 minutes (see *TMS Specific Procedures* section for more details). After the TMS study, all subjects were asked to record the eight syllables with their own voices with the same PsychToolbox script and set-up used in recording the original stimuli (see *Stimuli database* section above): this made it possible to subsequently measure the phonetic/articulatory distances between them and the heard stimuli.

In a second session (starting at least 60 days after the first TMS session), the same subjects were asked to complete the Multi-Dimensional Scaling (MDS) experiment (see *MDS Specific Procedures* section). Subjects rated the similarity of pairs of stimuli (i.e. pairs of audio files) containing all combinations of the original TMS experiment stimuli and the newly recorded TMS subjects' vocal productions. Subjects never judged their own voice recordings. The presentation was performed by means of a Psychtoolbox function running under Matlab environment (Mathworks, Inc.).

MDS Specific Procedures

We used the MDS data to obtain similarity ratings between experimental subjects' voice recordings (recorded in the first experimental session) and the speakers' voice recordings (the auditory stimuli in the TMS study). The aim of the MDS experiment was to build maps of subjective distances between the speech characteristics of the experimental subjects and the speech characteristics of the stimuli. Indeed, the stimuli were selected with the objective to create a highly variable auditory space to test the hypothesis of motor recruitment under demanding conditions, thus leading to the possibility to test how the subjects perceived this space and how they fitted in this space. In order to build these (dis)similarity maps, a new superset of stimuli was assembled. The superset included all the syllables used in the TMS experiment (the subset of the original recordings) and the subject's own vocal productions. Pairs were created following certain constraints: the elements within a pair were always from speakers of the same gender and always involved the same syllable. The number of pairs of stimuli presented to the subjects was given by the calculation of all possible pairs of syllables irrespective of order by (n(n-1)/2). This was adapted to the constraints described above, leading to the calculation of the number of possible pairs with male voices (10 male voices: 6 male voices obtained by recording the production of the 6 male subjects in the TMS experiment, plus 3 male voices obtained from the recordings of the speakers - used as stimuli in the TMS experiment - plus 1 filler voice, which was not presented during the TMS experiment, leading to a total of 45 pairs) plus number of possible pairs with female voices (10 female voices: 6 female voices from the recordings of the 6 female subjects in the TMS experiment, 3 female voices from the speakers and 1 female filler voice = 45 pairs), multiplied by the number of syllables (8 syllables) leading to a total of 720 pairs. Each subject was

presented with all pairs except those containing his/her own voice (72 pairs). These constraints resulted in 648 pairs. After listening to each pair, the subject reported the similarity by means of a visual-analogue scale (VAS) drawn on the computer screen, moving the mouse on a 100-step sliding cursor. Stimulus pair presentation was randomized for every subject and lasted approximately 90 minutes.

MDS Experiment Analysis

Multidimensional scaling was performed using the R statistical package (R Development Core Team 2008). First, a two-dimensional solution was obtained for each syllable separately; the subject ratings were averaged and used to obtain a perceived distance map across speakers for each syllable type (Figure 2). The cumulative sum of the first two eigenvalues divided by the sum of the absolute value of all eigenvalues for each solution was considered as a measure of goodness of fit of the two-dimensional solution, and validated two dimensions as being sufficient for the representation of the subjective ratings (we do not discuss the two dimensional analysis further, since the aim of the analysis was not to explore underlying dimensions but to obtain a simplified map of similarity distances). Maps of perceived distances were first obtained for each gender separately and were then merged together by centering the solution on the centroid given by averaging speakers' position, resulting in a common two-dimensional representation.

We then measured Euclidean distances between each subject's recording coordinates and each speaker's recording coordinates (in the common MDS two dimensional representation) to obtain a distance index. This index was used to investigate the relationship between the effect found in the TMS experiment and the subjective dissimilarity between subject-stimuli (speakers) individual characteristics. These

Euclidean distances were then included in a correlational analysis with RT-ratios data in R.

INSERT FIGURE 2 ABOUT HERE

Acoustic Distances Analysis

Acoustic distances were calculated between the subjects' own productions and the stimuli used during the TMS experiment, with the same method discussed above (cf. Figure 1). MFCC vectors were calculated and Principal Components Analysis was performed to reduce the dimensionality of the vectors. Formant estimations together with the principal components of the MFCC vectors were used to calculate Euclidean distances between each token. These distances were then used to investigate the relationship between TMS effects and the acoustic inter speaker distances by applying correlational analysis in R.

TMS Specific Procedures

TMS was delivered through a figure-eight coil (50 mm) and a Magstim Rapid stimulator (Magstim Co., Whitland, UK). For each subject, the left primary motor cortex was first localized following standard procedures (Rossini, 1994). Motor evoked potentials (MEPs) were recorded with adhesive Ag-AgCl electrodes (Digitimer 360 (Digitimer Ltd, Hertfordshire, UK); CED Power1401 (Cambridge Electronics, UK) acquisition board and Signal (version 4) software for data acquisition) from the right hand first dorsal interosseus muscle (FDI). The motor

threshold (MT) was identified as the lowest intensity of magnetic stimulation still able to elicit at least 50μV MEPs (Rossini 1994). Lips and tongue stimulation sites were localized on the basis of FDI localization, by using a combination of functional and probabilistic methods already used in our prior research to target tongue and lips motor representations (see D'Ausilio et al. 2009, 2012a). Specifically, Montreal Neurological Institute (MNI, Mazziotta at al. 2001) coordinates of the FDI, tongue and lips muscles were taken from previously published reports (lips: -56, -8, 46; tongue: -60, -10, 25; Pulvermüller et al. 2006; FDI: -37, -25, 58; Niyazov et al. 2005) and transformed in the 10-20 electroencephalography (EEG) system space (Steinsträter et al. in preparation, http://wwwneuro03.uni-muenster.de/ger/t2tconv/conv3d.html). We then computed differential coordinates between lips/tongue and FDI motor representations in the 10-20 space. This new reference system was therefore tailored on the individual cranial anatomy and was centered on the functionally defined FDI location.

The experiment consisted of two separate blocks with the same set of stimuli. In the two blocks, TMS (110% of the MT) was applied either to the lips or to the tongue motor area. The stimulation order was counterbalanced across subjects. Two single TMS pulses separated by 50 ms were delivered in TMS trials, starting 100 ms before speech onset. TMS was triggered by the pulse recorded in the second audio channel of the audio files.

TMS Experiment Analyses

All analyses were run using the R statistical package (R Development Core Team, 2008). Descriptive and diagnostic statistics established the validity of assumptions necessary for parametric statistics. RT values larger than 1500 ms and smaller than 50

ms were removed as outliers. Reaction times during no-TMS trials for dental syllables (766.35 ± 148. 56 ms) were not significantly different ($t(23)=-0.323$, $p=0.75$) from RTs for labial syllables (772.96 ± 180.79 ms), ruling out possible baseline differences in the discrimination time for the two classes of speech sounds considered. A possible confounding effect of TMS on reaction times (RTs) was ruled out by comparing average RTs in TMS trials (778.61 ± 158.23 ms) with those in no-TMS trials (769.66 ± 163.73 ms): no difference was found ($t(47)=0.724$, $p=0.47$). Subsequently, an index of the relative effect of TMS stimulation with respect to baseline performance was computed, by calculating the ratio between RTs during TMS and the RTs with no TMS for each stimulus. The ratio between RTs in TMS/no-TMS conditions has the advantage of giving an indication of the net effect of TMS on RTs while eliminating absolute differences between RTs, that can otherwise create large variability given the number of stimuli which subjects had to discriminate. Moreover, it allows us to simplify modeling the data by embedding in the dependent variable the effect related to the presence or absence of the magnetic pulse. This TMS/no-TMS ratio (which had a normal distribution according to Shapiro–Wilk normality test: $W = 0.9784$, $p = 0.51$) was then used as a dependent variable in repeated measures analysis of variance (rmANOVA). The factors included in the model were related to the point of articulation of the audio stimuli (2 levels: labials and dentals) and the TMS stimulation sites (2 levels: Lips_M1 and Tongue_M1). Accuracy (% of correct trials) was computed and analyzed as well.

Results

TMS Experiment Results

The direction of the TMS effect on behavioral measures (in this case, facilitation) depends on several factors, mainly represented by stimulation characteristics (frequency of pulses, timing with respect to stimulus) together with ongoing activity in the target area (Silvanto et al. 2008; Moliadze et al. 2003). Here, we adopted a stimulation procedure (see *TMS Specific Procedures* paragraph), which in our previous studies led to similar facilitation effects (D'Ausilio et al. 2009, 2011a).

A global ANOVA showed a significant interaction between the point of articulation of the speech stimuli and the site of TMS (condition*stimulation site: $F(1,11)=7.123$, $p=0.021$). Post-hoc comparison using Holm's adjustment for type 1 errors, revealed that the difference in the RT-ratio between labial and dental conditions was significant during tongue M1 stimulation ($p=0.006$) but not during Lips M1 stimulation ($p=0.19$). On the other hand, RT-ratios for the identification of dental stimuli showed a significant difference between the two stimulation sites ($p=0.049$), while labial stimuli showed almost as much difference ($p=0.052$). Generally speaking, RTs were shorter for dental syllables ([d],[t]) during stimulation of Tongue M1 (ratio RT TMS/no-TMS 0.94±0.2SEM) compared to Lips M1 stimulation (0.99±0.2). Labial syllables ([b],[p]) were more rapidly discriminated during Lips M1 stimulation (0.95±0.3) than during Tongue M1 stimulation (1.01±0.2). Therefore, stimulating the primary motor cortex of the effector involved in the production of a heard syllable speeds up the discrimination of that syllable in a specific and congruent manner (Figure 3). This result is in agreement with those in D'Ausilio et al. (2009).

That tongue motor area stimulation is more effective than that of the lips was already observed in our previous study (D'Ausilio et al. 2009). This might be due to several reasons: i) the larger variability in the tongue stimuli data-set (see *Stimuli Selection* paragraph); ii) the use of the first two formants (describing more tongue position) to

compute distances in the stimuli set; iii) the inclusion of vowels [a] and [i] which vary critically the tip of tongue position and only to a minor degree the lips during their pronunciation; iv) the greater overall variability in the role of the tongue in the production of phones (Chomsky and Halle 1968) and v) the differences in extension of the two somatotopic representations.

Responses were more accurate during no-TMS trials (93% responses) than during TMS trials (86% correct responses; $t(1151)=-6.734$, $p<0.001$) equally for each of the two stimulation sites (86% for both Lips M1 and Tongue M1) ($t(575)=0.246$, $p=0.81$). A slight difference was found between syllable types: the accuracy on dental syllable (84%) was lower than on labial ones (88%) ($t(575)=1.85$, $p=0.06$).

INSERT FIGURE 3 ABOUT HERE

Correlations between TMS data and distance measures

The MDS procedure measured the subjective distances between subjects' own voices and the stimuli in the TMS experiment. Here, the correlation analyses between TMS/NoTMS RT ratios of dental stimuli when the tongue M1 was stimulated yielded a significant relation ($r= 0.73$, $t(10) = 3.42$, p-value = 0.0066). Labial sounds when stimulating the Lips M1 showed no such ($r = -0.45$, $t(10) = -1.61$, p-value = 0.14) (see Figure 4, upper panels). The corresponding correlation analyses using objective similarity showed no significant relation with TMS/NoTMS RT ratio, neither for dental stimuli when the Tongue M1 was stimulated ($r= -0.38$, $t(10) = -1.31$, p-value = 0.22), nor for labial stimuli when the Lips M1 was stimulated ($r = 0.08$, $t(10) = 0.27$, p-value = 0.79) (Figure 4, lower panels).

INSERT FIGURE 4 ABOUT HERE

Discussion

The present study shows that the motor system can be critically involved in speech discrimination not only in conditions of degraded speech (D'Ausilio et al. 2009, 2012a, 2012b) but also when speech stimuli involve inter-speaker variability as the only source of "noise". This evidence rules out the possibility that the degree of motor involvement during speech discrimination was dependent on the difficulty of the task caused by acoustic interference in intelligibility.

Humans are extremely efficient in understanding speech despite distortions of many different kinds (Mattys et al. 2009). Distortions can be grouped according to their origin. External distortions are those related to environmental factors, such as background noise or signal filtering (Davis and Johnsrude 2003). Internal distortions are instead related to the specific characteristics of the speaker, such as vocal tract differences, rate, accent or style of speaking (Adank et al. 2009; Dupoux and Green 1997; Floccia et al. 2006).

Neuroimaging research on the processing of altered speech typically investigates stimuli with external background noise (Davis and Johnsrude 2003; Scott et al. 2004; Wong et al. 2009). These studies usually report noise-related activity in the bilateral Superior Temporal Gyrus (STG) and in the left Inferior Frontal Gyrus (IFG). While it is likely that the posterior STG is part of a pathway for processing comprehensible speech (Davis and Johnsrude 2003; Poldrack et al. 2001), the role of IFG activation is

less clear. A similar pattern of bilateral STG and left IFG activation has been shown while listening to speech pronounced with an unfamiliar accent (Adank et al. 2012a) and a further study showed that bilateral IFG activations were more noise-dependent: in contrast, accent processing elicited more activation of the left STG (Adank et al. 2012b).

In the present study, inter-speaker variability determined a causal contribution of the motor system to the syllable discrimination task. However, this is not in direct contradiction with the prior fMRI studies on the effect of accents because of the differences between the two situations (inter-speaker variability vs variable accents). In fact, Adank et al. (2012a, 2012b) used canonical and artificial accents, characterized by a set of vowel changes with a limited and predictable variability of the stimuli. On the contrary, in our study, speaker variability was natural, larger, and not systematic.

The present result is particularly relevant for two main reasons. First, it answers the concern that the motor system intervenes in speech perception only when the stimuli are acoustically degraded or the task difficulty is increased by noise (contra Hickok et al. 2011; Gernsbacher in Gallese et al. 2011). Our work shows, instead, that without signal degradation or noise and with a natural set of stimuli, the motor system can play a causal contribution to the discrimination task.

Second, our findings confirm the previous evidence that the listener's brain actively compares others' speech voices with its own motor production template. In the present study we asked to rate the similarity between each experimental stimulus and each TMS study participant's voice. In this way, by using a multidimensional scaling procedure, we could extract a subjective map of perceptual distances between listeners and speakers. Correlation between TMS effect on reaction times and these

distances showed that, the facilitation related to Tongue TMS was stronger for those participants whose mean perceptual distance was shorter from the TMS speakers with respect to the participants whose mean distance was greater, thus showing a gradient of motor recruitment from similar to dissimilar. In fact, perceived distance between two speakers can be proportional to the degree of motor recruitment during perception, as evidenced by the correlation analyses between TMS and MDS data. Accordingly, the interpretation that 'mirror' activities may depend on a weak modulatory role in perception of the motor system (Hickok et al. 2011), can hardly be reconciled with the somatotopically-related, specific motor activity, modulated by the perceptual self–others distances, shown in our present study.

The direct relationship between the amount of motor recruitment and inter-speaker subjective distance was limited to tongue-produced consonants. This may in part be explained by the relatively larger variance in the dental stimuli set. However, such differences in the stimuli set has been driven by other deeper characteristics of lips and tongue motor control. Tongue control has indeed been considered more complex from a biomechanical point of view. In fact, the tongue is believed to have 6 degrees of freedom whereas the lips only 3 (Beautemps et al. 2001). Moreover, on a more linguistic ground, the Distinctive Feature Theory (Chomsky and Halle 1968; Clements 1985) suggests that only two features are unique to lips positions whereas ten are unique to the tongue. In other words, tongue control is far more complex and thus reconstructing others' tongue configurations from an acoustic signal may prove computationally more difficult.

Therefore, the motor system may better show its contribution when discriminating sounds associated to complex motor control, as opposed to any other motorically simple discrimination. Here, we confirm that the motor contribution to speech

perception is stronger when the task poses some difficulties to the perceiver, ranging from the presence of noise in the signal (D'Ausilio et al. 2012a) to the presence of different speakers, as in the present experiment. In fact, the amount of motor recruitment may not be related to the signal to noise ratio in the auditory information but rather to the complexity of the motor control of that specific motor gesture. In line with this idea is the fact that the reaction time TMS/noTMS ratios and objective distances based on acoustic features did not correlate. This latter result suggests that the motor system does not take into account variance in the acoustic features of other's speech. Interestingly, subjective distances did not match the results of acoustic distances and in fact these measures did not correlate themselves ($r = 0.39$, $t(46) = 0.26$, p-value = 0.79). This is an interesting result by itself, since it shows that when subjects are asked to rate similarity between two speech segments they eventually map them in a space of features which are not necessarily acoustic only. Indeed, it is possible that participants were referring to a prototypical phonemic representation during the similarity judgment task, which, according to the theoretical framework of motor theories of speech perception, would be the participants' own production-to-perception mapping (Stevens and Halle 1967). By this view, the perception of others' speech features would be grounded on an internal representation, i.e. a template, formed by non-linear combination of features, which magnifies instances close to the prototype and represents distant speakers in a more coarse detail, thus giving rise to greater or smaller motor simulations.

Taken together all these findings are in favor of an active synthetic process (Kilner et al. 2004; Fadiga et al. 2006) more than a passive resonant mechanism. Active perception can be conceived as a model-based exploration of our environment. Although we don't perform explicit discrimination tasks in our daily activities, the

brain is always building inferences on the external world, actively rather than passively receiving data to classify. Indeed, coping with motor stimuli that are distant from our own motor repertoire may require an active search for specific motor invariants to enable classification. In agreement with such a predictive account, we have recently shown that the motor system activity anticipates incoming articulatory events by exploiting probability of phone occurrence and extracting subtle co-articulatory features from past signals (D'Ausilio et al. 2011b). These results are in line with recent computational models of motor control (Friston 2011) and mirror neuron mechanisms (Friston et al. 2011). According to these models both action and perception try to minimize surprise and the mirror-neuron mechanism may therefore implement a Bayesian-optimal perceptual processing of others' action.

Finally, the present results may have also some practical implication for automatic speech recognition (ASR). The issue of speech variability (as well as that of background noise) has always been a well-known problem for automatic speech recognition (Huang et al. 2001). As a matter of fact, variability is to be considered the main unresolved and weakest point in any computational model of speech processing. In our opinion, understanding why and how the brain copes with input variability and noise in such a robust manner could provide new insights for a new generation of automatic speech processing systems by incorporating the principle of active motor matching. In this direction are some attempts to acknowledge the beneficial role of articulatory features in improving phoneme/word classification (King et al. 2007). Very recently, our group has shown that the inclusion of articulatory knowledge during training of an artificial system for speech classification significantly increases classification accuracy, particularly in noisy conditions (Castellini et al. 2011). Epistemologically speaking, describing if, when and how, the variance of a specific

motor feature determines better accuracy in automatic speech recognition systems could be extremely informative on the relevant features used by the brain to solve similar tasks (Badino et al. In press).

Acknowledgments

This work has been supported by EU grants SIEMPRE and POETICON++. The authors declare no competing financial interests.

References

Adank P, Davis MH, Hagoort P. 2012a. Neural dissociation in processing noise and accent in spoken language comprehension. Neuropsychologia. 50:77–84.

Adank P, Evans BG, Stuart-Smith J, Scott SK. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. Journal of experimental psychology Human perception and performance. 35:520–529.

Adank P, Noordzij ML, Hagoort P. 2012b. The role of planum temporale in processing accent variation in spoken language comprehension. Human brain mapping. 33:360–372.

Badino L, Ausilio AD, Fadiga L, Metta G. In press. Computational modeling and validation of the motor contribution to speech perception. Topics in Cognitive Science.

Beautemps D, Badin P, Bailly G. 2001. Linear degrees of freedom in speech production: analysis of cineradio- and labio-film data and articulatory-acoustic modeling. Journal of the Acoustical Society of America 109:2165-80.

Beyer K, Goldstein J, Ramakrishnan R, Shaft U. 1999. When Is "Nearest Neighbor" Meaningful? Lecture Notes in Computer Science 1540: 217-235.

Binder JR, Liebenthal E, Possing ET, Medler D a, Ward BD. 2004. Neural correlates of sensory and decision processes in auditory object identification. Nature neuroscience. 7:295–301.

Callan D, Callan A, Gamez M, Sato M, Kawato M. 2010. Premotor cortex mediates perceptual performance. NeuroImage. 51:844–858.

Callan DE, Jones J a, Callan AM, Akahane-Yamada R. 2004. Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. NeuroImage. 22:1182–1194.

Castellini C, Badino L, Metta G, Sandini G, Tavella M, Grimaldi M, Fadiga L. 2011. The use of phonetic motor invariants can improve automatic phoneme discrimination. PloS one. 6:e24055.

Chomsky N, Halle M. 1968. The sound pattern of English. New York: Harper and Row.

Clements GN. 1985. The Geometry of Phonological Features. Phonology Yearbook 2:225-252.

Davis MH, Johnsrude IS. 2003. Hierarchical processing in spoken language comprehension. The Journal of neuroscience : the official journal of the Society for Neuroscience. 23:3423–3431.

Diehl RL, Lotto AJ, Holt LL. 2004. Speech perception. Annual review of psychology. 55:149–179.

Dupoux E, Green K. 1997. Perceptual adjustment to highly compressed speech: effects of talker and rate changes. Journal of experimental psychology Human perception and performance. 23:914–927.

D'Ausilio A, Pulvermüller F, Salmas P, Bufalari I, Begliomini C, Fadiga L. 2009. The motor somatotopy of speech perception. Current biology. 19:381–385.

D’Ausilio A, Bufalari I, Salmas P, Busan P, Fadiga L. 2011a. Vocal pitch discrimination in the motor system. Brain and language. 118:9–14.

D’Ausilio A, Jarmolowska J, Busan P, Bufalari I, Craighero L. 2011b. Tongue corticospinal modulation during attended verbal stimuli: priming and coarticulation effects. Neuropsychologia. 49:3670–3676.

D’Ausilio A, Bufalari I, Salmas P, Fadiga L. 2012a. The role of the motor system in discriminating normal and degraded speech sounds. Cortex. 48:882–887.

D'Ausilio A, Craighero L, Fadiga L. 2012b. The contribution of the frontal lobe to the perception of speech. Journal of Neurolinguistics. 25:328–335.

Fadiga L, Craighero L, Buccino G, Rizzolatti G. 2002. Speech listening specifically modulates the excitability of tongue muscles : a TMS study. European Journal of Neuroscience. 15:399–402.

Fadiga L, Craighero L, Destro MF, Finos L, Cotillon-Williams N, Smith AT, Castiello U. 2006. Language in shadow. Social neuroscience. 1:77–89.

Floccia C, Goslin J, Girard F, Konopczynski G. 2006. Does a regional accent perturb speech processing? Journal of experimental psychology Human perception and performance. 32:1276–1293.

Flury B. 1988. Common Principal Components and Related Multivariate Models. New York: Wiley.

Fowler C. 1986. An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics. 14:3–28.

Friston K. 2011. What is optimal about motor control? Neuron. 72:488–498.

Friston K, Mattout J, Kilner J. 2011. Action understanding and active inference. Biological cybernetics. 104:137–160.

Gallese V, Gernsbacher M a., Heyes C, Hickok G, Iacoboni M. 2011. Mirror Neuron Forum. Perspectives on Psychological Science. 6:369–407.

Hickok G, Houde J, Rong F. 2011. Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. Neuron. 69:407–422.

Huang X, Acero A, Hon H. 2001. Spoken language processing. Upper Saddle River, New Jersey: Prentice Hall PTR.

Kilner JM, Vargas C, Duval S, Blakemore S-J, Sirigu A. 2004. Motor activation prior to observation of a predicted movement. Nature neuroscience. 7:1299–1301.

King S, Frankel J, Livescu K, McDermott E, Richmond K, Wester M. 2007. Speech production knowledge in automatic speech recognition. The Journal of the Acoustical Society of America. 121:723–742.

Liberman A, Cooper F. 1967. Perception of the speech code. Psychological Review. 74:431–461.

Londei A, D'Ausilio A, Basso D, Sestieri C, Gratta C Del, Romani G-L, Belardinelli MO. 2010. Sensory-motor brain network connectivity for speech comprehension. Human brain mapping. 31:567–580.

Mattys SL, Brooks J, Cooke M. 2009. Recognizing speech under a processing load: dissociating energetic from informational factors. Cognitive psychology. 59:203–243.

Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Le Goualher G, Boomsma D, Cannon T, Kawashima R, Mazoyer B. 2001. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences. 356:1293–1322.

Meister IG, Wilson SM, Deblieck C, Wu AD, Iacoboni M. 2007. The essential role of premotor cortex in speech perception. Current biology. 17:1692–1696.

Moliadze V, Zhao Y, Eysel U, Funke K. 2003. Effect of transcranial magnetic stimulation on single-unit activity in the cat primary visual cortex. The Journal of physiology. 553:665–679.

Möttönen R, Watkins KE. 2009. Motor representations of articulators contribute to categorical perception of speech sounds. The Journal of neuroscience : the official journal of the Society for Neuroscience. 29:9819–9825.

Niyazov DM, Butler a J, Kadah YM, Epstein CM, Hu XP. 2005. Functional magnetic resonance imaging and transcranial magnetic stimulation: effects of motor imagery, movement and coil orientation. Clinical neurophysiology : official

journal of the International Federation of Clinical Neurophysiology. 116:1601–1610.

Phillips PC, Arnold SJ. 1999. Hierarchical comparison of genetic variance-covariance matrices. I. using the Flury hierarchy. Evolution. 53(5):1506-1515.

Poldrack RA, Temple E, Protopapas A, Nagarajan S, Tallal P, Merzenich M, Gabrieli JD. 2001. Relations between the neural bases of dynamic auditory processing and phonological processing: evidence from fMRI. Journal of cognitive neuroscience. 13:687–697.

Pulvermüller F, Huss M, Kherif F, Moscoso del Prado Martin F, Hauk O, Shtyrov Y. 2006. Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences of the United States of America. 103:7865–7870.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rossini P. 1994. Non-invasive electrical and magnetic stimulation of the brain, spinal cord and roots: basic principles and procedures for routine clinical application. Report. Electroenceph Clin Neurophysiol. 91:79–92.

Sato M, Tremblay P, Gracco VL. 2009. A mediating role of the premotor cortex in phoneme segmentation. Brain and language. 111:1–7.

Scott SK, Rosen S, Wickham L, Wise RJS. 2004. A positron emission tomography study of the neural basis of informational and energetic masking effects in

speech perception. The Journal of the Acoustical Society of America. 115:813–821.

Shahin AJ, Bishop CW, Miller LM. 2009. Neural mechanisms for illusory filling-in of degraded speech. NeuroImage. 44:1133–1143.

Silvanto J, Muggleton N, Walsh V. 2008. State-dependency in brain stimulation studies of perception and cognition. Trends in cognitive sciences. 12:447–454.

Stevens KN, Halle M. 1967. Remarks on analysis by synthesis and distinctive features. In: Wathen-Dunn W. Models for the Perception of Speech and Visual Form. Cambridge: M.I.T. press. p88-102.

Watkins K. 2003. Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia. 41:989–994.

Wilson SM, Saygin AP, Sereno MI, Iacoboni M. 2004. Listening to speech activates motor areas involved in speech production. Nature neuroscience. 7:701–702.

Wong PCM, Jin JX, Gunasekera GM, Abel R, Lee ER, Dhar S. 2009. Aging and cortical mechanisms of speech perception in noise. Neuropsychologia. 47:693–703.

Captions

Figure 1: Stimulus selection procedure

Example of segments selection for one given syllable. The upper panel shows the configuration of distances obtained by calculating f1 and f2, whereas the lower panel is relative to the distances derived by the principal component of mel-frequencies cepstral coefficients (MFCC) method, (only the first 2 of 15 principal components are plotted), The grey rectangles represent all segments from all speakers for the [di] syllable, the black dot is the starting segment selected for the iterative maximization algorithm, the black rectangles represent the other segments chosen by the integration of the two distance-based methods.

Figure 2: Subjective inter-speaker distance calculation

Example of subjective inter-speaker distance calculation for one given syllable. The MDS procedures enable the projection of subjective distances on an arbitrary 2 dimensional space. Asterisks, bounded by black lines, represent speakers' vocal productions used as stimuli in the TMS study. Circles show the relative positions of TMS subjects' own vocal productions. Grey lines connect one representative subject to each TMS stimulus, thus representing the computation of distances between that subject and stimuli presented in the TMS study.

Figure 3: TMS results

Bar plots represent mean reaction times (TMS/no-TMS ratio) for the syllable identification task. In the abscissa are shown the two sites of TMS stimulation (Lips

M1 and Tongue M1). Dark grey bars represent the dental syllables ([da], [di], [ta], [ti]) and in light grey, the labial syllables ([pa], [pi], [ba], [bi]). Corrected p-values are shown for each post-hoc comparison.

Figure 4: Correlation analyses

Correlation analyses between TMS effect on RTs (ratio between RTs in TMS and no-TMS trials) and the distance measures. Upper panel (A) shows the correlation with the subjective distances (MDS procedure), whereas the lower one (B) shows the correlation with the objective acoustic measures. On the left side of both panels are represented the correlations during the stimulation of the Lips M1 and the labial syllables. On the right side instead are shown the correlations with the stimulation of Tongue M1 and dental stimuli.