TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2014) 1–15
Copyright © 2014 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12095

Computational Validation of the Motor Contribution to Speech Perception

Leonardo Badino,[a] Alessandro D'Ausilio,[a] Luciano Fadiga,[a,b] Giorgio Metta[a,c]

[a]*RBCS – Robotics, Brain and Cognitive Sciences Department, IIT – Istituto Italiano di Tecnologia*
[b]*DSBTA – Section of Human Physiology, University of Ferrara*
[c]*DIST – Dipartimento di Informatica, Sistemistica, Telematica, University of Genova*

Received 30 March 2012; received in revised form 4 October 2012; accepted 29 November 2012

Abstract

Action perception and recognition are core abilities fundamental for human social interaction. A parieto-frontal network (the mirror neuron system) matches visually presented biological motion information onto observers' motor representations. This process of matching the actions of others onto our own sensorimotor repertoire is thought to be important for action recognition, providing a non-mediated "motor perception" based on a bidirectional flow of information along the mirror parieto-frontal circuits. State-of-the-art machine learning strategies for hand action identification have shown better performances when sensorimotor data, as opposed to visual information only, are available during learning. As speech is a particular type of action (with acoustic targets), it is expected to activate a mirror neuron mechanism. Indeed, in speech perception, motor centers have been shown to be causally involved in the discrimination of speech sounds. In this paper, we review recent neurophysiological and machine learning-based studies showing (a) the specific contribution of the motor system to speech perception and (b) that automatic phone recognition is significantly improved when motor data are used during training of classifiers (as opposed to learning from purely auditory data).

Keywords: Motor theory of speech perception; Transcranial magnetic stimulation; Automatic speech recognition; Machine learning

1. Introduction

Speech production is a special motor ability, as it requires fine motor control and temporal coordination of different articulators. Similarly, the ability to perceive speech

Correspondence should be sent to Leonardo Badino, IIT – Fondazione Istituto Italiano di Tecnologia, RBCS – Robotics, Brain and Cognitive Sciences Department, Via Morego, 30, 16163 Genova, Italy. E-mail: leonardo.Badino@iit.it

can be viewed as the ability to recognize this special type of motor act. The idea that articulatory goals are important for perception was proposed by the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), the theory of analysis by synthesis (Stevens & Halle, 1967), and the direct realist theory (Fowler, 1986). Generally speaking, all these theories embrace a constructivist approach, suggesting that speech perception is mediated by constraints imposed by a sensorimotor model. The direct realist theory (Fowler, 1986) proposes that, although there are no acoustic features that invariantly specify the units of speech, there are invariant properties in sensory stimulation that unambiguously and directly specify the articulatory gestures, responsible for production. This model, in fact, does not require any inferential process. According to this approach, what we perceive is not sensory in nature but directly relates to the articulatory gesture (Bever & Poeppel, 2010; Pulvermüller & Fadiga, 2010).

A motor contribution to the solution of perceptual tasks is also in line with neurophysiological research. Monkey area F5, a ventral premotor area, contains neurons responding to the execution of a given specific goal-directed action (i.e., grasping, manipulating, tearing, or holding; Rizzolatti et al., 1988). In addition, several F5 neurons also show complex visual responses (visuomotor neurons). Two categories of these visuomotor neurons are present in area F5: canonical and mirror neurons. Mirror neurons discharge both when the monkey executes and when it observes another individual making the same action in front of it (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). The most likely interpretation for the visual discharge of mirror neurons is that it performs a mapping from visual biological motions to the observer's motor repertoire, facilitating the classification of action and thus others' action understanding.

Recent research in the domain of speech perception has been driven by this recent resurgence of interest on the role of the motor system in perception. However, several other theories suggest that categorical speech perception is based on a purely sensory analysis of the incoming stimuli (Diehl, Lotto, & Holt, 2004). Several lines of evidence are often cited to support this possibility. One of these lines reasons that animals that cannot produce speech (e.g., chinchillas) can, nonetheless, develop categorical speech perception in a manner characteristic of human listeners (Kuhl & Miller, 1975). In this seminal study, chinchillas were trained to differently respond to t/and/d/consonant–vowel syllables showing generalization to novel instances, and also the same phonetic perception boundaries of English-speaking adults. There is no doubt that these animals can perform the task adequately. It is to be noted that in computational terms, this is an extremely simple problem as the data set is large, variability is low, and the classes to be discriminated are few. In fact, this task can be performed by using a purely acoustic feature extraction. Current automatic speech recognition (ASR) systems perform extremely well in such simplified and unnatural scenarios, but their recognition accuracy is often far from being acceptable when speech is largely variable, that is, noisy, distorted, spontaneous, and with large inter-speaker and speaking style variability (Benzeghiba et al., 2007). Our belief is that the motor system plays a central role, especially in these cases, and it is crucial during development. To support this hypothesis, we study the motor contribution to speech perception from two different and complementary perspectives, a neurobiological perspective and a computational one.

From the neurobiological perspective, we carried out transcranial magnetic stimulation (TMS) experiments to verify (a) whether the activity of the motor system has a causal influence on speech perception; (b) when such activity is more critical for speech perception; and (c) the mechanisms that allow such activity to affect speech perception.

From a computational perspective, we try to validate the actual utility of motor information in automatic speech classification/recognition[1] tasks comparing "motor-informed" classifiers/recognizers with state-of-the-art classifiers/recognizers that only rely on speech acoustics. The computational perspective can provide answers to questions that neurophysiological techniques alone cannot answer (or would have a very hard time to answer) and provide hints on why the brain activates motor areas during speech perception. Examples of such questions are as follows: Why should the brain use motor information to recognize speech sounds? What are the limits of a brain that only uses auditory information (like in the case of chinchillas)?

Following the synthetic approach as proposed, for instance, by Pfeifer, Lungarella, and Sporns (2008), we do not aim at a very detailed structural model of the cortical speech perception processes and the related brain areas, but rather we seek an appropriate abstraction by considering the type of data that are available to the brain and the functional properties of their interconnections. The goodness of such abstraction is then measured in terms of functional similarity with humans and thus of accuracy in perception tasks.

All perception tasks we address concern the perception of phonemes. The contribution of (pragmatic, semantic, phonotactic, etc.) context to speech perception is not considered.

2. Motor contribution to speech perception from a neurobiological perspective

Research on the role of the motor system in speech perception has used a behavioral approach for several decades (Galantucci, Fowler, & Turvey, 2006). In the past 20 years, with the advent of novel neurophysiological and neuroimaging techniques, the involvement of the motor system can be measured directly. In fact, several studies have found motor and premotor activations while listening to speech sounds (Binder, Liebenthal, Possing, Medler, & Ward, 2004; Callan, Jones, Callan, & Akahane-Yamada, 2004; Callan et al., 2010; Londei et al., 2010; Pulvermüller et al., 2006; Skipper, Nusbaum, & Small, 2005; Wilson, Saygin, Sereno, & Iacoboni, 2004). Other studies using transcranial magnetic stimulation (TMS) showed specific cortico-spinal facilitation while passive listening to speech sounds (D'Ausilio, Jarmolowska, et al., 2011; Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Murakami, Restle, & Ziemann, 2011; Roy, Craighero, Fabbri-Destro, & Fadiga, 2008; Watkins, Strafella, & Paus, 2003). However, one of the strong predictions of motor theories at large is that activities in motor centers are functionally necessary for perception. In this sense, TMS on motor centers should alter subjects' performance in perceptual tasks. This has been shown in a variety of tasks and different TMS protocols

(D'Ausilio, Bufalari, et al., 2011; D'Ausilio et al., 2009; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Möttönen & Watkins, 2009; Sato, Tremblay, & Gracco, 2009).

Two recent studies applied repetitive TMS (rTMS) to the premotor region activated both during speech production and speech discrimination (Wilson et al., 2004). Stimulation impaired classification of degraded acoustic stimuli (Meister et al., 2007) and phoneme discrimination when the task required a relatively high degree of processing load (Sato et al., 2009). Moreover, it has been shown that rTMS applied over lips primary motor cortex alters the discrimination and identification of verbal stimuli containing different proportions of motorically discordant syllables (Möttönen & Watkins, 2009).

In our studies, we instead applied online focal TMS to the lips or tongue motor regions. Magnetic stimulation facilitated the discrimination of tongue- or lips-produced phonemes ([b] and [p] vs. [d] and [t]), thus showing effects causally associated with the motor system (D'Ausilio et al., 2009). Furthermore, the double dissociation between stimulation site and place of articulation of the verbal stimuli suggests that the effect is somatotopically specific. This is a strong support to the claim that dissociable cortical representations are causally implicated in the processing and classification of speech sounds produced with different articulators. However, we also showed that the motor system might be more critical when performing demanding phonological tasks (D'Ausilio, Craighero, & Fadiga, in press; see left panel of Fig. 1). In fact, our results are in line with several other studies showing that anterior language areas might be recruited for sensory decisions and completion during sub-optimal listening conditions (Binder et al., 2004; Boatman & Miglioretti, 2005; Moineau, Dronkers, & Bates, 2005).

Future lines of research will need to investigate the exact conditions that favor the contribution of the motor system in these discrimination tasks. Furthermore, an interesting and open question is if individual articulatory differences explain the amount of motor recruitment during speech perception. In fact, motor activations might be scaled for the distances between listener and speakers' relative motor and/or auditory representations.

Proponents of purely acoustic theories have highlighted other sources of empirical data against a causal contribution of the motor system in speech perception tasks (Diehl et al., 2004). For example, perception of speech sounds is not abolished in patients who have severely impaired speech production due to chronic stroke (Naeser, Palumbo, Helm-Estabrooks, Stiassny-Eder, & Albert, 1989; Weller, 1993) or in individuals who never acquired the ability to speak due to congenital disease or prelingual brain damage (Bishop, Brown, & Robson, 1990; Christen et al., 2000). However, testing of speech comprehension deficits was not the main goal of these studies. In fact, the fairly good receptive functions displayed by these patients were often anecdotally reported or tested with clinical scales, possibly measuring general language comprehension rather than speech discrimination abilities. More important, clinical language tests could dissociate respect to speech discrimination. In fact, language comprehension in natural scenarios and even clinical settings usually does not require complete phone recognition because contextual information, such as sematic and syntactic knowledge, actually enables strong top-down predictions. Therefore, in our opinion, these studies offer important insights for discussion, but they cannot offer a good case to reject motor theories, at least at the

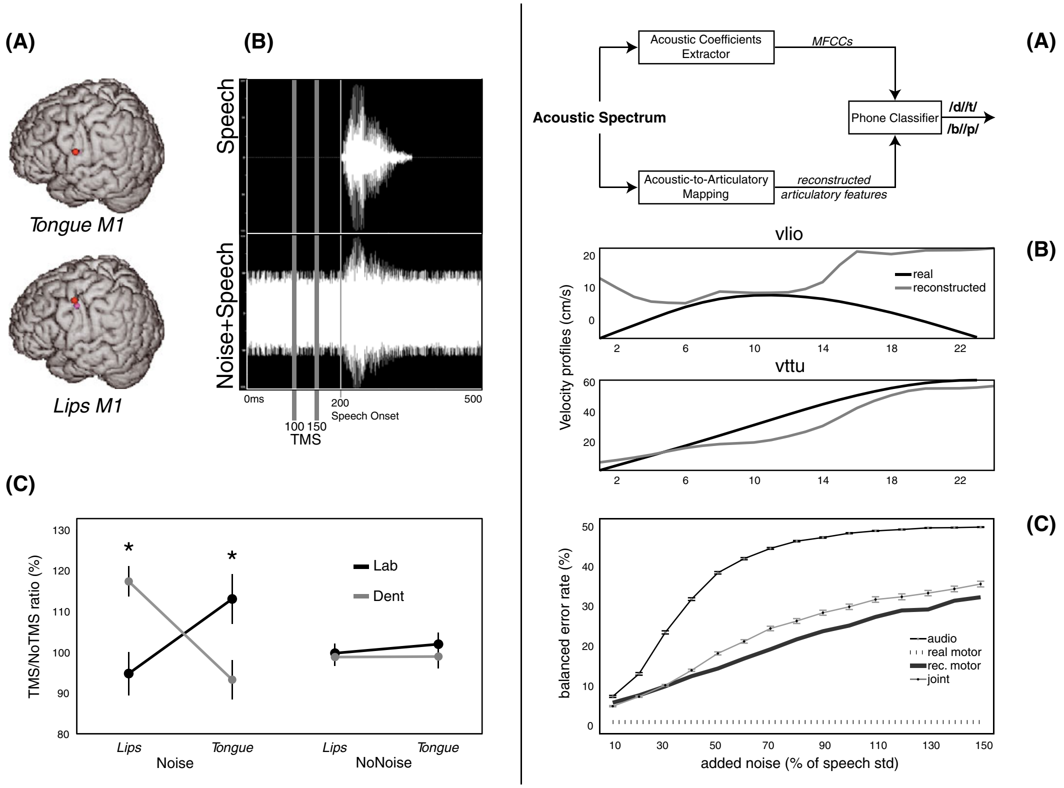

Fig. 1. Neurobiological and computational approaches and results. The left panel shows the methods and results stemming from our research on the neurobiological basis of speech perception. The right part of the figure shows the parallel computational research on the same topic. The left panel shows the cortical sites where TMS was applied in our studies (A; tongue and lips motor cortex) as well as the timing of TMS pulses and the kind of stimuli (B; clean or embedded in white noise). The graph in the lower left panel summarizes the results of the magnetic stimulation of the motor system during our speech discrimination task (C;/p/,/b/ ,/d/and/t/). TMS facilitated subjects' responses only when stimuli were embedded in white noise (D'Ausilio et al., 2009, in press). The right panel shows the flow of processes in our computational studies (A). Real and AMM-reconstructed velocity of lips distance (vlio) and velocity of tongue tip – upper teeth distance (vttu) for subject 6 uttering the/t/ (B). Notice the apparent gap in the quality of the reconstruction, favoring in this case the labiodental trajectory (vttu). The last graph (C) shows the balanced error rate in /t/,/d/vs./p/ ,/b/classification at different Signal/Noise ratios (Castellini et al., 2011).

speech level. Furthermore, another study often cited to contrast motor theories, used a well-designed word comprehension test in individuals who have acute and complete deactivation of the left or right hemisphere due to left carotid artery injection of sodium amobarbital (Wada procedure) (Hickok et al., 2008). Although a clear speech arrest can be induced for the left hemisphere deactivation, this functional ablation is not specific to motor centers and leaves the right motor system intact. More important, these subjects were presented with well-designed behavioral tests targeting again language comprehension rather than the speech discrimination ability.

Summing up, there is a large amount of research demonstrating the activation of motor centers during speech perception tasks, as well as very recent evidence of its causal contribution (TMS interference paradigms). On the other hand, critical issues often affect research demonstrating intact receptive abilities following chronic or temporary functional lesion of motor centers. These problems are typically related to the poor brain structure specificity or to the use of suboptimal tasks testing wider linguistic abilities.

3. Motor contribution to speech perception from a computational perspective

One of the strongest motivations to build machines that recognize articulatory gestures to aid speech recognition is the presence of many phenomena observed in speech, for example, coarticulation effects, for which a simple purely acoustic description has still to be found. These phenomena can be easily and compactly modeled when considering the gestures of the vocal tract articulators. Additionally, phonetic target gestures, that is, the motor gestures that are targeted to utter phonemes, the typical basic units in ASR, are far more invariant to speaker's vocal tract, speaking style, and, obviously, to environmental noise than the acoustic features of phonemes.

However, during speech recognition, the articulatory gestures need to be recovered from speech and one may wonder whether the problem of recovering articulatory gestures is as hard as the problem of recognizing words directly from acoustics only. Some of the limitations of current state-of-the-art ASR systems can give us clues on (a) general (i.e., that do not depend on the technique used to perform ASR) problems that any system (being it an artificial or a biological one) that attempts to recognize speech has to tackle when relying on acoustics only and (b) shed some light on why and how articulatory gestures could provide solutions that would be far more complex and/or difficult to be found if gestures were not recovered from acoustics.

From now on, we will assume that the subwords, that is, the basic units of an ASR system, are phonemes. However, our discussion also applies to other linguistic units, such as, for example, syllables. Note that basic linguistic units such as phonemes and syllables could be replaced by purely articulatory units in a more "extreme" motor-based ASR system as proposed by articulatory phonology (Browman & Goldstein, 1992).

3.1. Speech recognition without motor information

In ASR, speech is typically represented as a sequence of discrete and disjoint acoustic states (this is usually referred to as "beads-on-a-string" approach Ostendorf, 1999), whereas speech is actually the result of overlapping and continuous movements of the vocal tract articulators that move to reach their targets. The assumption that states are disjoint imposes that the dynamics of speech can only be crudely modeled as the structural similarities between states are unknown (because there are no shared parts between them). As an example, in ASR systems based on Hidden Markov Models (HMMs), the most widely used framework for ASR (see, e.g., Huang, Acero, & Hon, 2001 for basics of Hidden Markov Models–based ASR), coarticulation is handled by creating a model (a parametric probability density function) for each possible phonetic context of each phoneme. That results in a very large number of context-dependent models, which cause data sparseness issues. Note that we would have the same explosion of models even when using longer span units (e.g., syllables) as proposed by purely auditory theories (e.g., Massaro, 1972). The number of models can be reduced by means of data-clustering methods that make some implicit use of speech production knowledge. However, such implicit use of motor information does not avoid the models (actually, the clusters gathering them) to be largely dependent on speaking style, environmental noise, etc., that is, all other causes of speech variability that are not due to coarticulation.

All those limitations can be largely reduced and phenomena like coarticulation can be easily modeled when we move to the articulatory space. Each phoneme, now defined as a (spatio-temporal) target configuration of the critical articulators (i.e., the articulators that are necessary to utter that given phoneme), has properties (subtargets) that are shared with the other phonemes. That allows a compact representation of speech and models of transitions between phonemes that are "informed" of their structure. Additionally, such an articulatory structure may allow articulation-based prediction on the next acoustic events; that is, the articulatory information acts as a prior for auditory perception (as well as pragmatics, semantics, etc.). Coarticulation can be modeled as asynchrony between articulators where asynchrony is due to anticipation or preservation of adjacent phones that force articulators to reach their own target at different time points. The behavior of few parameters (the articulators), rather than a large number of parametric models, is sufficient to model the effects of the phonetic context.

It cannot be excluded that a pure acoustic structure, as effective as the articulatory structure, might be found in the future, but why should we only search for this structure at the surface level of speech and ignore the causes of speech? In fact, most structures of speech proposed in the ASR literature are based on knowledge of the speech production processes (Deng, Yu, & Acero, 2006; Livescu, Glass, & Bilmes, 2003) or on directly measured articulatory information (Markov, Dang, & Nakamura, 2006).

Articulatory data may be useful for ASR even when its structure is ignored. Some studies (e.g., Wrench & Richmond, 2000; Zlokarnik, 1995) have shown that even in an approach where the potential of articulatory information is only partly exploited by simply adding measured articulator positions to the set of acoustic features (typically Mel-filtered Cepstra coefficients, MFCCs, which are computed through a linear cosine transform of a log-power spectrum on a perceptual scale, the Mel scale, of frequency) of a standard, "beads-on-a-string"-based, ASR system, they produce increased accuracy rates. In our studies on the contribution of measured articulatory data to phone classification and recognition, we have so far investigated this "non-structured" approach.

3.2. *Recovering motor information from acoustics*

In a real scenario, where only acoustic information is available to the recognizer during recognition, any approach that uses measured articulatory information (during training) requires the articulatory information to be recovered from acoustics. This recovering is performed by an acoustic-to-articulatory mapping (AAM) function, typically constructed by learning on simultaneous recordings of speech and articulatory movements. If the AAM function is applied to recover the positions of the articulators (plus variations over time) from the spectrum, the recovered data can serve as complementary features to the standard acoustic features (e.g., MFCCs). Even when the AAM function is applied to recover the articulatory data from the standard acoustic features, an increased recognition accuracy is still possible because the AAM acts as a transformation function on the acoustic features space that carries information about the speech production process (with all the advantages that come from it).

The main requirement for any approach that relies on AAM functions is that those functions must be good enough. A few studies that use measured articulatory data for ASR are mainly concerned with the fact that the AAM problem does not have a unique solution. Identical sounds can be produced by posing the articulators in a range of different positions (Lindblom, Lubker, & Gay, 1979) and, in some cases, the conditional probability of the position of an articulator given the acoustic evidence is even multi-modal (Roweiss, 1999).

While methods have been proposed to address this ill-posedness of AAM (Richmond, King, & Taylor, 2003; Toda, Black, & Tokuda, 2007), a mechanism accounting for the different importance of the articulators in the realization of a given phoneme is lacking. Forcing to recover the gestures of non-critical articulators as accurately as those of the critical ones may result in forcing to recover irrelevant, and potentially noisy, information, to the detriment of the relevant ones. This might be one of the reasons why reconstructed articulatory information in ASR has not been supported by consistent strong empirical evidence yet (see the comprehensive review by King et al., 2007).

In Castellini et al., 2011 (see right panel of Fig. 1), we largely reduced the technical difficulties implied by AAM mentioned above by addressing a phonetic binary classification problem (Italian /p/, /b/ vs. /t/, /d/), which is the “computational” counterpart of our TMS study (D’Ausilio et al., 2009). Note that phone classification is strictly related to the estimation of phone probabilities, given the acoustic evidence. The phone posterior probability estimation is the only ASR “subtask” that can exploit articulatory information (if we ignore the articulatory structure).

The only articulatory data (recorded with an electromagnetic articulograph, see Grimaldi et al., 2008, for details on the corpus used) taken into account were the data that distinguish the two classes, namely the velocities and accelerations of the Euclidian distances from the two lips and from the tongue tip to the upper teeth (positions where excluded for technical reasons). Results showed that a classifier based on real articulatory features dramatically outperforms all other classifiers (one based on MFFCs, one on recovered articulatory features, and one on both of them) in any task. In general, when the classification task becomes more difficult (i.e., the ratio between variability in the training data and variability in the testing data decreases), the reconstructed articulatory features lead to significant improvements with respect to the baseline, either when combined with the audio features or alone. The utility of the recovered articulatory features is striking when the classifiers are tested on noisy speech.

It must be noted that simply adding new information in a classifier usually results in an increased "capacity" of the classifier (where the maximum "capacity" is the ability to learn any training dataset without error), but not necessarily in a better "generalization" (where the best generalization is achieved when the classifier makes no errors on the testing dataset, which is completely disjoint from the training dataset). In Castellini et al. (2011), we show that when adding motor information, the generalization of the classifiers increases especially when achieving a good generalization becomes more critical because the training data set (and/or the speech variability within it) available is reduced, then the motor information becomes more relevant.

Qin and Carreira-Perpiñán (2007) showed (on a single speaker data) that although the non-uniqueness of AAM is normal in human speech, most of the time the vocal tract has a unique configuration when producing a given phone. On the other hand, non-linearity seems to be a much more important aspect of AAM. The primacy of non-linearity is supported by the fact that multi-layer perceptrons, that is, feed-forward neural networks, which can properly handle AAM non-linearity but cannot handle AAM non-uniqueness, are one of the best performing machine learning strategies for AAM (Mitra, Nam, Espy-Wilson, Saltzman, & Goldstein, 2010).

In Badino, Canevari, Fadiga, and Metta (2012) we tested the utility of articulatory information on a full speaker-dependent phone recognition task, by using a hybrid Deep Neural Network-Hidden Markov Model system. Deep neural networks were used to 1) perform AAM and 2) estimate the phone posterior probabilities, given the acoustic and (recovered) articulatory evidence. With the term deep neural networks (DNNs), we mean feed-forward neural networks whose parameters are first "pre-trained" using unsupervised training of deep belief networks (Hinton, Osindero, & Teh, 2006) and subsequently fine-tuned using backpropagation. In other words, DNNs are an improved version of feed-forward networks that exploits the knowledge of the statistical properties of the input domain (i.e., P(X)) to effectively guide the search for input–output relations (i.e., P(Y|X)). To handle the temporal dynamics of speech, DNNs can be combined with Hidden Markov Models. DNN-HMM systems are the state-of-the-art machine learning strategy in automatic phone recognition (Mohamed, Dahl, & Hinton, 2012).

When combining recovered articulatory features with MFCCs, the phone recognition error reduction, with respect to the same DNN-HMM phone recognizer only using MFCCs, was 16% on the MOCHA-TIMIT corpus (Wrench, 2000). We believe that such a remarkable result is mainly due to the fact that DNNs exploit the relevant articulatory information for phone posterior estimation better than other methods typically used in HMM-based systems for phone posterior estimation (i.e., Gaussian mixtures).

The actual utility of articulatory information for speech recognition may depend on the kind of representation of the articulatory domain used. In most of the previous work on the use of measured articulatory information for speech recognition, the articulatory domain consists of a set of independent articulator trajectories (plus their first and second derivative). Such domain can be explicitly transformed in a new domain that would become the new target of AAM.

The motivation behind a transformation of the original articulatory domain is that an improved and more useful recovering of the articulatory data may be achieved if we use representations of the articulatory data that compactly encode the behavior of each articulator in relation to other articulators or parts of the vocal tract (e.g., upper teeth). In fact, in articulatory phonetics, the articulatory features (e.g., the consonant places of articulation) that distinguish phones almost always refer to the relative position of a critical articulator, with respect to another critical articulator or a specific part of the vocal tract.

Obviously, any feature set that is extracted from the positional feature set can contain at most the same amount of information contained in the original feature set. However, by trying to recover features that capture the most frequent dependencies (potentially phone-distinctive) between parts of the vocal tract, we may remove information that is irrelevant for the discrimination task and retain most of the information that is needed. Removing irrelevant information is important both to reconstruct the articulatory information (because we are not forced to learn an AAM function to reconstruct irrelevant and, most probably, noisy information) and to effectively use the reconstructed articulatory information to estimate the phone posterior probabilities (because, again, we ignore irrelevant and potentially noisy information). Additionally, by using this new kind of representations, we might collapse all articulatory data points that produce an identical sound in one single small cluster, thus "reducing" the ill-posedness of the AAM problem.

The *vocal tract* features proposed by articulatory phonology (Browman & Goldstein, 1992) are an interesting example of an alternative representation of the articulatory domain. They are relative features with interesting properties, for example, a smaller non-uniqueness than independent (to each other) articulator positions, and significantly improve word recognition in noisy speech conditions (Mitra, Nam, Espy-Wilson, Saltzman, & Goldstein, 2012).

In Badino et al. (2012), we proposed an alternative, data-driven, representation of the articulatory domain. The new representation was obtained by transformation of the original domain (consisting of articulator trajectories plus their first and second derivatives) through deep auto-encoder networks (Hinton & Salakhutdinov, 2006). Deep auto-encoder networks are a particular type of DNNs that transform the original input domain into a new "encoded" domain that should capture the strongest dependencies between features of the original domain. In our experiments, the (recovered) transformation of the articulatory domain produced a phone recognition accuracy increase (with respect to the non-transformed articulatory features) on the MOCHA-TIMIT corpus. Although the resulting accuracy increase is small, it certainly encourages further work on new data-driven representations of the articulatory domain.

Other future directions will include (a) a more "structured" use of articulatory information where the dynamics of the articulators will be explicitly modeled; (b) creation of corpora containing spontaneous conversational speech; and (c) study of strategies to exploit motor information in "ecological scenarios" where the phone recognition task is speaker independent and motor information is available to one single speaker (while only the acoustics of other speaker is available during training).

4. General discussion

We have presented an approach to study speech perception that searches for neurophysiological evidence for the hypothesis that motor information contributes to speech perception and test such hypothesis on automatic speech recognition tasks. The hypothesis is supported both by TMS studies and by results showing significantly increased automatic phone recognition accuracy when measured articulatory data recovered from acoustics are combined with acoustic features.

Both TMS and computational studies show that the relevance of motor information increases when the difficulty of the perception task increases. TMS of the motor centers significantly affects speech perception when the perceived speech is noisy, whereas in automatic phone recognition tasks, the positive impact of reconstructed motor features on recognition accuracy increases when the difficulty of the classification task increases. In real life and in most of real ASR applications, recognizing noisy speech and, more in general, speech that has not been heard beforehand (because of new kinds of noise, of channel distortions, of new speakers, etc.) is the norm, whereas recognizing clean, clearly articulated, and "well-known" speech is the exception.

The results of our experiments cannot completely reject the hypothesis that the motor system is not necessary for speech perception. Indeed, all the necessary empirical information (which we separate from the prior information provided by pragmatics, semantics, phonotactics, etc.) for speech recognition is in the acoustic domain, from which motor information is recovered. Thus, humans might still accurately recognize speech even if they extracted all necessary information from acoustics without relying on any motor information. However, it would not be clear why they should not exploit the benefits (ranging from structuring of speech to extraction of discriminative acoustic features) brought by knowledge of the speech production process when they are available for free. Summing up, our position suggests that information, in perfect listening scenarios, is highly redundant. So much redundant that several different strategies could be employed, requiring different degrees of motor intervention. However, there are two critical points to be discussed here. The first is that "perfect listening conditions" almost never apply in real life (imagine inter-speaker differences as a source of unavoidable noise). Secondly, and more computationally relevant, the use of motor knowledge enables representations of the incoming stimulus where the acoustic properties of phonological units (e.g., phonemes) are more invariant; that is, less dependent on, for example, phonetic context, or environmental noise. Therefore, here we are not suggesting a limited role of the motor system in speech perception; rather, we suggest that any speech recognition can be performed in different ways with varying costs. If necessary (in case of brain lesion or TMS stimulation), the recognition can be performed on acoustic only data; however, this is not the most efficient nor computationally realistic manner of performing such a task.

Acknowledgments

The authors acknowledge the support of the European Commission project CONTACT (grant agreement NEST-5010), Poeticon++ (grant agreement FP7-ICT-288382) and SIEMPRE (grant agreement FP7-IST-250026)

Note

1. In a phone classification task, the phone boundaries are given and the classifier has to assign a phonetic class to each speech segment, whereas in a phone recognition task, utterances are not segmented by phone boundaries and the recognizer has to identify the correct sequence of phonemes associated with the input utterance. Occasionally, we will use the term *recognition* to indicate both classification and recognition when there is need to distinguish the two concepts.

References

Badino, L., Canevari, C., Fadiga, L., & Metta, G. (2012). *Deep-level Acoustic-to-Articulatory Mapping for DBN-HMM based phone recognition*. Miami, FL: IEEE Workshop on Spoken Language Technology.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*, 763–786.

Bever, T. G., & Poeppel, D. (2010). Analysis by Synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics*, *4*(2–3), 174–200.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, *7*, 295–301.

Bishop, D. V., Brown, B. B., & Robson, J. (1990). The relationship between phoneme discrimination, speech production, and language comprehension in cerebral palsied individuals. *Journal of Speech and Hearing Research*, *33*, 210–219.

Boatman, D. F., & Miglioretti, D. L. (2005). Cortical sites critical for speech discrimination in normal and impaired listeners. *Journal of Neuroscience*, *25*, 5475–5480.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica*, *49*(3–4), 155–180.

Callan, D. E., Jones, J. A., Callan, A. M., & 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*(9), e24055. doi:10.1371/journal.pone.0024055.

Christen, H. J., Hanefeld, F., Kruse, E., Imhauser, S., Ernst, J. P., & Finkenstaedt, M. (2000). Foix-Chavany-Marie (anterior operculum) syndrome in childhood: A reappraisal of Worster-Drought syndrome. *Developmental Medicine & Child Neurology*, *42*(2), 122–132.

D'Ausilio, A., Bufalari, I., Salmas, P., Busan, P., & Fadiga, L. (2011). Vocal pitch discrimination in the motor system. *Brain and Language*, *118*(1–2), 9–14.

D'Ausilio, A., Craighero, L., & Fadiga, L. (2012). The contribution of the frontal lobe to the perception of speech. *Journal of Neurolinguistics*, *25*(5), 328–335.

D'Ausilio, A., Jarmolowska, J., Busan, P., Bufalari, I., & Craighero, L. (2011). Tongue corticospinal modulation during attended verbal stimuli: priming and coarticulation effects. *Neuropsychologia.*, *49*(13), 3670–3676.

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.

Deng, L., Yu, D., & Acero, A. (2006). Structured speech modeling. *IEEE Transaction on Audio Speech and Language*, *5*(4), 1492–1504.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.

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.

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

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin Review*, *13*, 361–377.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609.

Grimaldi, M., Gili Fivela, B., Sigona, F., Tavella, M., Fitzpatrick, P., & Graighero, L., et al. (2008). *New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph*. Proceedings of LangTech 2008. Rome, Italy.

Hickok, G., Okada, K., Barr, W., Pa, J., Rogalsky, C., Donnelly, K., et al. (2008). Bilateral capacity for speech sound processing in auditory comprehension: Evidence from Wada procedures. *Brain and Language*, *107*(3), 179–184.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing*. Englewood Cliffs, NJ: Prentice-Hall.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, *121*(2), 723–742.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*, 69–72.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, *7*, 146–161.

Livescu, K., Glass, J., & Bilmes, J. (2003). Hidden feature modeling for speech recognition using dynamic Bayesian networks. *Proceedings of Eurospeech, Geneva, Switzerland*, *4*, 2529–2532.

Londei, A., D'Ausilio, A., Basso, D., Sestieri, C., Del Gratta, C., Romani, G. L., et al. (2010). Sensory-motor brain network connectivity for speech comprehension. *Human Brain Mapping*, *31*, 567–580.

Markov, K., Dang, J., & Nakamura, S. (2006). Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. *Speech Communication*, *48*, 161–175.

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, *79*(2), 124–145.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, *17*, 1692–1696.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2010). Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE J of Selected Topics in Signal Processing*, *4*(6), 1027–1045.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2012). Recognizing articulatory gestures from speech for robust speech recognition. *Journal of the Acoustical Society of America*, *131*(3), 2270–2287.

Mohamed, A., Dahl, G. E., & Hinton, G. E. (2012). Acoustic Modeling using Deep Belief Networks. *IEEE Trans. On Audio, Speech, and Language Processing*, *20*(1).

Moineau, S., Dronkers, N. F., & Bates, E. (2005). Exploring the processing continuum of single-word comprehension in aphasia. *Journal of Speech, Language and Hearing Research*, *48*, 884–896.

Möttönen, R., & Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *Journal of Neuroscience*, *29*, 9819–9825.

Murakami, T., Restle, J., & Ziemann, U. (2011). Observation-execution matching and action inhibition in human primary motor cortex during viewing of speech-related lip movements or listening to speech. *Neuropsychologia.*, *49*, 2045–2054.

Naeser, M. A., Palumbo, C. L., Helm-Estabrooks, N., Stiassny-Eder, D., & Albert, M. L. (1989). Severe nonfluency in aphasia: Role of the medical subcallosal fasciculus and other white matter pathways in recovery of spontaneous speech. *Brain*, *112*, 1–38.

Ostendorf, M. (1999). Moving beyond the "beads-on-a-string" model of speech. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (Keystone, CO), vol. 1, pp. 79–83.

Pfeifer, R., Lungarella, M., & Sporns, O. (2008) The synthetic approach to embodied cognition: A primer. In O. Calvo & A. Gomila (Eds.), *Handbook of cognitive science: An embodied approach* (pp. 121–137). Amsterdam: Elsevier.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360.

Pulvermüller, F., Huss, M., Kherif, F., del Prado, Moscoso, 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.

Qin, C., & Carreira-Perpiñán, M. A. (2007). *An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping*. Interspeech, Antwerp, Belgium: Proc.

Richmond, K., King, S., & Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, *17*(2), 153–172.

Rizzolatti, G., et al. (1988). Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. *Experimental Brain Research*, *71*, 491–507.

Roweiss, S. (1999). *Data driven production models for speech processing*. PhD thesis. Pasadena, CA: California Institute of Technology.

Roy, A. C., Craighero, L., Fabbri-Destro, M., & Fadiga, L. (2008). Phonological and lexical motor facilitation during speech listening: A transcranial magnetic stimulation study. *Journal of Physiology Paris*, *102*, 101–105.

Sato, M., Tremblay, P., & Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain & Language*, *111*, 1–7.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage*, *25*, 76–89.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Walthen-Dunn (Ed.), *Models for the perception of speech and visual form*. Cambridge, MA: MIT Press.

Toda, T., Black, A., & Tokuda, K. (2007). Statistical mapping between articulatory movements and Acoustic spectrum using a gaussian mixture model. *Speech Communication*, *50*(3), 215–222.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*, 989–994.

Weller, M. (1993). Anterior opercular cortex lesions cause dissociated lower cranial nerve palsies and anarthria but no aphasia: Foix-Chavany-Marie syndrome and “automatic voluntary dissociation” revisited. *Journal of Neurology*, *240*(4), 199–208.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*, 701–702.

Wrench, A. A. (2000). A multichannel articulatory database and its application for automatic speech recognition. *Proceedings 5th Seminar of Speech Production*, Kloster Seeon, Germany.

Wrench, A. A., & Richmond, K. (2000). Continuous speech recognition using articulatory data. *Proceedings of the International Conference on Spoken Language Processing*, pp. 145–148.

Zlokarnik, I. (1995). Adding articulatory features to acoustic features for automatic speech recognition. *Journal of the Acoustical Society of America*, *97*(2), 3246.