Demodulation, mirror neurons and audiovisual perception nullify the motor theory

Hartmut Traunmüller
Department of Linguistics, University of Stockholm

Abstract

According to the Motor Theory of Speech Perception (MTSP), listeners perceive speech by way of the articulatory gestures they would perform themselves in producing a similar signal. The theory postulates a module that allows extracting gestural information from the signal. The gestures constitute the event perceived.

According to the Modulation Theory (MDT), speech is modulated voice. Listeners perceive it by demodulating the signal. The properties of the voice convey non-linguistic information while the linguistically coded information is conveyed by its modulation. The modulation pattern constitutes the linguistic event perceived.

The theories agree in requiring a linkage or mapping between perception and production. According to MDT, phonetically labeled links between exteroception and proprioception (mirror and echo neurons) are established in the brain during speech acquisition. The set of links embodies the knowledge of the relation. While MDT describes the device that MTSP would need in order to be implemented, it makes it redundant to recruit the motor system. Demodulation is also necessary in speechreading and in order to perceive sign language, when a face or body is 'modulated' instead of a voice. In audiovisual speech perception, there are two percepts: a normally dominant vocal one and a gestural one that does not need to agree with it. MTSP knows of only one of these. It is concluded that all the specific claims of MTSP are false while MDT rests on 'first principles'.

Two theories

A theory based on articulatory gestures

For almost 50 years now, the Motor Theory of Speech Perception, MTSP (Liberman, Cooper, Harris & MacNeilage, 1962; Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967) has been more widely propagated than any less controversial alternative. Instead of the sound, MTSP considers the articulatory gestures used in the production of speech to be perceived. The percept is represented by the neuromotor commands that the listener would activate himself if he were producing the perceived utterance. According to the revised version (Liberman & Mattingly, 1985), the gestures *intended* by the speaker constitute the 'object of perception'.

MTSP was suggested subsequent to the observation that the acoustic properties of the same speech sounds were not invariant but context dependent. Liberman's group was not aware of the investigations by Menzerath and de Lacerda (1933), which had shown a similar lack of invariance in the articulatory gestures and that these cannot be fully recovered from the sound.

The theory postulates the existence of a cognitive module that is specialized for speech and allows listeners to derive the articulatory activity that corresponds to a speech signal. The approach remained computationally empty since the functionality of the module was not specified sufficiently. It has, however, been stated that an analysis-by-synthesis procedure would be needed. Models of speech perception based on this kind of procedure have been elaborated in some detail (Halle & Stevens, 1962), but they represent a differently focused approach.

In short, MTSP claims that

1. speech processing is modular,
2. perceiving speech is perceiving gestures and
3. the motor system is recruited for perceiving speech.

While the experimental research done by Liberman's group has been seminal in other respects, MTSP has not really been productive and has failed to be more widely accepted within phonetics. It has found more positive acceptance outside phonetics, and arguments advanced in favor of some of its tenets have sometimes been

drawn from perception of gesture rather than speech (Galantucci, Fowler & Turvey, 2006).

Recently, MTSP has attracted some attention in research concerned with the mirror-neuron system (Rizzolatti & Craighero, 2004; Fadiga, Craighero, Buccino & Rizzolatti, 2002) and with functional magnetic resonance imaging (fMRI) of cortical activity (Pulvermüller, Huss, Kherif, Moscoso del Prado Martín, Hauk & Shtyrov, 2006; Skipper, Nusbaum & Small, 2007). Investigations of this kind have disclosed a linkage between perception and production of speech at the cortical level. Such a linkage is also required by MTSP. It is, then, tempting to proclaim the theory as confirmed by these observations. However, the observations of Pulvermüller et al. (2006) argue against modularity of speech processes and those of Skipper et al. (2007) show substantial activity in speech motor areas to be evoked during *visual* rather than auditory perception of speech.

In the following, it will be argued that a linkage by auditory mirror neurons ('echo neurons') makes it redundant to recruit the motor system in normal speech perception. This conclusion emerges upon considering how the linkage is established during speech acquisition, which is outside the scope of MTSP.

Even when only speech perception in experienced listeners is considered, MTSP suffers from serious shortcomings. To these belongs

- the insufficiency of the theory for modeling speech perception,
- its failure to cope with non-linguistic variation due to perspective and between-speaker differences in organs of speech and, last not least,
- the fact that an ability to speak is not necessary in order to perceive speech.

While an analysis-by-synthesis approach allows avoiding these shortcomings, this also requires substituting an abstract model of organs of speech for the listener's own.

A theory based on demodulation

The Modulation Theory, MDT (Traunmüller, 1994; 2000), which is a 'demodulation theory' as far as speech perception is concerned, resulted from a line of research in which the interplay between the linguistic and the non-linguistic aspects of human speech was in focus. It has recently been further developed in order to capture the linkage between perception, production and representation of speech in the brains of speaker/listeners (Traunmüller, ms). Its underlying principles are applicable not only to communication by speech but also by gesture.

MDT is founded on an analysis of the nature and equivalence conditions of speech signals. Its essentials can be expressed in four propositions:

1. A speech signal is basically the result of a process in which a carrier – the speaker's voice – has been modulated by phono-articulatory gestures.
2. The auditory properties of speech signals with the same linguistically informative quality deviate from those of an unmodulated carrier essentially in the same specific way.
3. For speech perception, this implies that a demodulation is necessary in order to recover and separate the different kinds of information.
4. Speakers associate the auditory modulation patterns of speech signals with the kinesthetically and haptically sensible properties of their own speech gestures with the same linguistic phonetic quality.

The carrier can be thought of as a linguistically neutral vocalization, an [ə], produced in a relaxed way, to the extent to which expressive factors allow this.

Prop. 2 defines the modulation pattern of the voice as the linguistically informative event and 'object' of production as well as perception.

In speech perception, listeners 'tune in' to the speaker's voice, and they evaluate the deviations of the auditory properties of the signal from those they expect of an [ə] with the same para- and extralinguistic quality (same attitude, emotion, mouth and perspective).

Prop. 4 links the perception of the linguistically informative quality of speech, which is described by a sequential trace in a multidimensional space of distinctive properties (features or dimensions), with the somatosensory feedback that allows speakers to *control* their speech production appropriately if they strive to modulate their voice in a desired linguistic way. Thus, it links exteroception with proprioception and not with efferent phases of production.

MDT invokes sensory processes in production, but it does not invoke motor processes in perception. As for perception, it is primarily concerned with the process that MTSP presup-

poses to occur without specifying it in any detail. MDT describes this process as a demodulation and provides the foundation for modeling it.

The role of echo neurons

The linkage described by prop. 4 may be realized by echo neurons that can be excited alike either by the detected modulation of a voice (exteroceptive) or by somatosensory properties (interoceptive) of the same phonetic unit. Such an echo neuron in fact *equates* these cases. Each echo neuron represents a distinctive dimension or feature, but there may be many that represent variants of the same. Speech segments are represented in the simultaneous activity of the set of echo neurons.

Infants can be assumed to establish echo neurons of this kind by babbling, when the sound is heard coincidently with the gesture being felt. Quantal phenomena in modulation patterns and discontinuities in the somatotopic representation of articulatory gestures (cf. Pulvermüller et al., 2006) will lead to a natural clustering of echo neurons. During the course of speech acquisition, the set of echo neurons will further be partitioned when language specific phonetic labels become attached to the echo neurons as a result of cumulative experience with the language.

In estimating the similarity of different stimuli, subjects can be assumed to take all the associations evoked by the stimuli into account, i.e., all the linkages of the excited echo neurons. Therefore, the labeling has the effect of increasing the perceptual difference between stimuli that fall into different categories. This has been observed in investigations of "categorical perception" and provides also an explanation for the "perceptual magnet effect" (Kuhl, 1991).

While the input on the exteroceptive side is dominated by the auditory modulation of a voice, there is also an entrance for the visual modulation of a face, which can be quite important for easily visible features (cf. McGurk & MacDonald, 1976). There is an entrance from two sensory modalities also on the interoceptive side: kinesthetic and haptic.

After the stage of speech acquisition, all the implicit knowledge a language user has about the relation between exteroception and proprioception of the same phonetic units and strings is permanently laid down in the set of echo neurons and in memorized sequential traces. As soon as an echo neuron is activated, all this implicit knowledge becomes highlighted. Recruiting motor processes in perception will not bring in anything in addition to that. While echo neuron activity due to perception may be detectable in cortical regions associated with speech production, this does not mean that motor processes have been invoked, but just that an equivalent articulation has been indicated.

The motor system may be involved actively in attempts to perceive an unfamiliar distinction, e.g. in a second language, since this is likely to involve processes typical of speech acquisition. An fMRI study by Callan, Jones, Callan and Akahane-Yamada (2004) showed higher activity in Japanese as compared with native English subjects exposed to [l] and [r]. A study with a large and varied set of consonants in an [ɑ_ɑ] context by Wilson and Iacoboni (2006) revealed that the increase in neural activity covaries with the *producibility* of non-native consonants only in *auditory* areas. MTSP would rather lead one to expect this in motor areas.

Prop. 4 does not exclude the possibility of additional linkages between perception and production. The neuromotor commands activated in production are affected by expressive factors (attitudes and emotions) but could be linked with a stage in speech perception where the expressive quality has not yet been factored out and definite phonetic labels have not yet been attached. Such a link may work in the other direction in "shadowing" (cf. Shockley, Sabadini & Fowler, 2004) and give rise to motor evoked potentials (Fadiga et al., 2002). The latter could, however, also arise without such an additional linkage, which is more likely to link *visual* perception of speech gestures with articulation. This is suggested not only by the observation that shadowers whenever possible focus on the shadowed speaker's lips, but also more directly by fMRI data obtained by Skipper et al. (2007).

Speech and gesture

In principle, MDT is applicable to speechreading and to sign language as well, and to the imitation of any bodily gestures and postures. In these cases, there is a face, body or skeleton that is 'modulated' instead of a voice. In order to make the propositions that describe MDT apply to gestural language, they need to be translated, substituting "visual properties of gestures" for "auditory properties of speech signals" etc.

When the perspectival and organic variation in gestures has been factored out, the modulation describes the gestures directly, except for their invisible parts. With this reservation, the

distinction between perceived and produced gestures appears to vanish. This trivializes the second claim of MTSP and has evoked hopes of finding arguments in favor of the others in the perception of gesture instead of speech (Galantucci et al., 2006). It may explain why speech motor activity is more easily evoked by *visual,* i.e., truly gestural perception of speech (Skipper et al., 2007). However, even in gesture perception, demodulation is indispensable since perspectival and organic variation has to be factored out and, in the case of linguistic communication, expressive variation as well. Masked parts of a gesture or a vocal modulation pattern will normally be filled in, in the prejudiced way in which perception works, by knowledge derived from previous experience, rather than by recruiting the motor system.

Speechreading is perceiving gesture according to MDT as well as MTSP, but only MDT considers ordinary speech perception to be different in that it is perceiving vocal modulations. In audiovisual speech perception, both theories suggest that the optic signal may have an effect, but MTSP knows only of one percept, which is gestural in nature, while MDT allows for a gestural percept in *addition* to a vocal one.

The results of recent experiments with incongruent audiovisual stimuli (Traunmüller, 2006) have, indeed, shown a gestural percept to coexist with a separate speech percept. In these experiments, subjects had to tell either what they heard or what they saw and to rate the distinctive features involved. In both tasks (listening and speechreading), the results showed an influence of the non-attended modality, but the strength of this influence varied between the tasks in a feature specific way.

While certain cortical areas in which echo neurons are likely to terminate are involved in perception as well as in production of speech, and a neural linkage between perception and production is fundamental for the faculty of speech, the three specific claims of MTSP must now all be considered as false while MDT rests firmly on 'first principles'.

Acknowledgement

This contribution was supported by grant 421-2004-2345 from the Swedish Research Council.

References

Callan DE, Jones JA, Callan AM & Akahab-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.

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.

Galantucci B, Fowler CA & Turvey MT (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*: 361-377.

Halle M & Stevens K (1962). Speech recognition: A model and a program for research. *IEEE Transactions on Inforamtion Theory, 8*: 155-159.

Kuhl PK (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics, 50*: 93-107.

Liberman AM, Cooper FS, Harris KS & MacNeilage PF (1962). A motor theory of speech perception. In the Proc. of the *Speech Communication Seminar*, Stockholm, Vol. II, Paper D3.

Liberman AM, Cooper FS, Shankweiler DP & Studdert-Kennedy M (1967). Perception of the speech code. *Psychological Review, 74*: 431-461.

Liberman AM & Mattingly IG (1985). The motor theory of speech perception revised. *Cognition, 21*: 1-36.

McGurk H & MacDonald J (1976). Hearing lips and seeing voices. *Nature, 264*: 746-748.

Menzerath P & de Lacerda A (1933). *Koartikulation, Steuerung und Lautabgrenzung: Eine experimentelle Untersuchung.* Berlin: Dümmler.

Pulvermüller F, Huss M, Kherif F, Moscoso del Prado Martín F, Hauk O & Shtyrov Y (2006). Motor cortex maps articulatory features of speech sounds. *P. Natl. Acad. Sci. USA, 103*: 7865-7870.

Rizzolatti G & Craighero L (2004). The mirror-neuron system. *Annu. Rev. Neurosci., 27*: 169-192.

Shockley K, Sabadini L & Fowler C (2004). Imitation in shadowing words. *Perception and Psychophysics, 66*: 422-429.

Skipper JI, Wassenhove V van, Nusbaum HC & Small SL (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, advance access, doi:10.1093/cercor/bhl147.

Traunmüller H (1994). Conventional, biological and environmental factors in speech communication: A modulation theory. *Phonetica, 51*: 170-183.

Traunmüller H (2000). "Evidence for demodulation in speech perception" *Proc. of the 6th ICSLP, vol III*: 790-793.

Traunmüller H (2006). Cross-modal interactions in visual as opposed to auditory perception of vowels. *Working Papers, 52*: 137-140. Dept. of Linguistics & Phonetics, Lund Univ.

Traunmüller H (ms). *Speech considered as modulated voice*. Dept. of Linguistics, Stockholm Univ.

Wilson SM & Iacoboni M (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage, 33*: 316-325.