Journal of Neurolinguistics 25 (2012) 355–381

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling

Auditory representations and phonological illusions: A linguist's perspective on the neuropsychological bases of speech perception

Andrea Calabrese

Department of Linguistics, University of Connecticut, 337 Mansfield Road, Storrs, CT 06269, United States

ARTICLE INFO

Article history:
Received 13 July 2009
Received in revised form 20 January 2011
Accepted 29 March 2011

Keywords:
Phonological illusions
Top-down and bottom-up speech perception
Analysis-by-synthesis
Echoic memory
Phonology
Distinctive features
Auditory representations
Syllable structure
Mirror neurons

ABSTRACT

This paper argues that speech perception includes grammatical—in particular phonological—computations implemented by an analysis-by-synthesis component (Halle & Stevens, 1962) which analyzes linguistic material by synthesizing it anew. Analysis-by-synthesis, however, is not always required in perception but only when the listener wants to be certain that the words or morphemes identified in the input signal correspond to those intended by the speaker who produced the signal (= parity requirements, see Liberman, 1996; Liberman & Whalen, 2000). As we will see, in some situation analysis-by-synthesis may generate 'phonological' illusions. A central assumption is that the representations of words or morphemes in perception involve distinctive features and are formally structured into syllables. Two perceptual modes are needed: phonetic and phonemic perception. In phonemic perception only contrastive aspects of sounds, i.e., the aspect of sounds associated with meaning differences, are searched for. In phonetic perception both contrastive and non-contrastive aspects of sounds are identified. The phenomenon of phonological 'deafening' will be shown to follow from phonemic perception.

The paper also argues that the perception system must include an echoic memory component (Neisser (1967)) where faithful auditory representations of acoustic inputs can be stored. This echoic memory is part of a bottom-up system of perception dedicated to the collection and storage of the acoustic signal.

The paper ends with the discussion of some hypotheses (and related questions) on the neural bases of perceptual processes and

E-mail address: andrea.calabrese@uconn.edu.

0911-6044/$ – see front matter
doi:10.1016/j.jneuroling.2011.03.005

representations. A brief assessment of the role of mirror neurons in perception is given here.

1. Introduction

This article has its origin in a project on loanword phonology started a few years ago. Working on that project, I discovered the importance that speech perception plays in accounting for the adaptations found in loanwords and for my own peace of mind I had to study this phenomenon more in depth. In this paper I discuss some of the opinions, beliefs and hypotheses that I developed in doing this research. They are the thoughts of a theoretical linguist who in his intellectual inquiry tries to make sense of the vast phonetic, psychological and neurological literature on this argument in the light of what he knows and believes about language.[1]

Recent literature has shown that when listeners experience non-native, unfamiliar sounds and sound combinations, they may have 'illusions' in which the perception of these sounds or sound combinations is distorted. These 'illusions' indicate that perception must involve processes which are part of a perceptual system that is dedicated—top-down—to analysis, identification, and recognition of linguistic information in the acoustic inputs. The important point is that these processes involve grammatical—in particular phonological—computations of the same type as those found in production (Brown, 1998, 2000; Matthews & Brown, 2004). It is my belief that the best way to account for the intervention of these grammatical computations in perception is to assume that perception includes an analysis-by-synthesis component (Halle & Stevens, 1962) which analyzes linguistic material by synthesizing it anew through grammatical computations.

However, if perception involves top-down grammatical processes, we should expect that sounds that are illicit from a grammatical point of view, i.e., the non-native, unfamiliar sounds and sound combinations mentioned above, could never be perceived correctly insofar as they would always be distorted in 'perceptual illusions'. Therefore, a learner would never be able to learn them properly. This is contrary to the common human experience of foreign sounds. It is a fact that despite possible perceptual distortions, learners have access to foreign sounds even though they cannot articulate and recognize them, and will eventually learn them correctly by constructing articulatory representations that approximate their acoustic reality.

The acoustic reality of the speech input must therefore also be accessible in perception. To account for this fact, I hypothesize that a fundamental role in perception is played by echoic memory (Neisser, 1967) where faithful auditory representations of acustic inputs are stored. This echoic memory is part of a bottom-up system of perception dedicated to the collection and storage of the acoustic signal.

A fundamental theme throughout this article will be the issue of how we perceive and learn unfamiliar sound configurations, in particular words containing unfamiliar sound configurations, and how these sound configurations are adjusted during this process. It will include a brief discussion of the issue of phonological deafening in language learning.

It is important to say that in my work as a linguist I subscribe to the realistic approach to language advocated by Bromberger and Halle (1992, 1997, 2000) (see also Halle, 2002; Calabrese, 2005). According to this approach, "phonology is about concrete mental events and states that occur in real time, real space, have causes, have effects, are finite in number." (Bromberger & Halle, 2000: 21). This is in contrast to what Bromberger and Halle (2000: 99) call linguistic Platonism according to which phonology—and linguistics—is about abstract, non-spatio-temporal objects. Platonistic approaches to language, which treat language only as pure mathematical computation and disregard the fact that

[1] Many of the ideas discussed here were elaborated for my original work on loanword phonology (Calabrese, 2009). I later discovered that one of these ideas, specifically the importance of the notion of analysis-by-synthesis in speech perception, was one of the centerpieces of the model of speech perception developed in Poeppel et al. (2008) (see also Bever & Poeppel, 2010; Poeppel & Idsardi, 2010; Poeppel & Monahan, 2010). Reading these papers made me rethink various aspects of my ideas and analysis. The present paper owes much to them.

language has a concrete bodily base, are quite common in linguistics. They lead to abstract ideas that have only a remote relation with the actual reality of language as produced in real time through a complex interaction between body and brain. In the realistic approach, the reality of language involves our concrete acts of speech performed by our limited bodies and brains, and the theory of phonology—and linguistics—must be built on this reality. Linguistic computations must be executed in the brain in real time. Under such an approach, the primitives and the operations that are involved in constitutive subroutines of linguistic computations must be such that they can plausibly be executed by assemblies of neurons in real time. The linguist must therefore find ways to bridge the gap between mind and brain and to connect the abstract models of linguistic theory concerning the mental computation occurring in language to the study of the organization and functioning of the neural circuits performing such computations. My hope is that this article may provide some small hints on how to achieve these goals in the case of speech perception.

2. Phonological illusions

Peperkamp and Dupoux (2003) review psycholinguistic evidence showing that all aspects of non-native—'unfamiliar'—phonological structure, including segments, suprasegments, and syllable phonotactics, may be systematically distorted during speech perception. A striking case showing such distortion is the perception of illusory vowels within consonantal clusters by Japanese speakers. Japanese listeners perceive illusory epenthetic vowels in sequences of segments that do not fit the syllable structure of their native languages. In a series of behavioral experiments, Dupoux et al. (Dehaene-Lambertz, Dupoux, & Gout, 2000; Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Dupoux, Pallier, Kakehi, & Mehler, 2001; Dupoux, Peperkamp, & Sebastián-Gallés, 2001) compared Japanese listeners with French listeners in their perception of consonant clusters. For instance, Dupoux et al. (1999), in an off- line phoneme detection task (Experiment 1) used a series of six items created from naturally produced nonce words (e.g., [abuno], [akumo], [ebuzo], [egudo], etc.) by gradually reducing the duration of the vowel [u] down to 0 ms. The participants were asked to answer whether each item they heard included the sound [u]. Japanese listeners, unlike French listeners, overwhelmingly reported that the vowel was present at all levels of vowel length. Strikingly, this was the case seventy percent of the time even when the vowel had been completely removed (i.e., the zero ms condition). The French participants, on the other hand, judged that the vowel was absent in the no-vowel condition about 90% of the time and that a vowel was present in only 50% of the intermediate cases. These results were confirmed in other experiments, which have led Dupoux et al. to conclude that listeners 'invent' illusory vowels in perception to accommodate sequences of segments that are illicit in their L1.

Similar results were obtained by Kabak and Idsardi (2007) for Korean: Korean listeners perceive illusory vowels within consonantal clusters that are otherwise illicit according to the syllable structure of their language. Kabak and Idsardi's article is important because it demonstrates that the illusory vowels observed in Korean are due to grammatical restrictions on syllable structure. In particular, they show that epenthesis occurs only in the case of sequences such as *ct, cm* that are excluded because *c* cannot be found in coda position (see sect. 6) and not in the case of sequences such as *km* or *ln* that, although never found in the native phonology of Korean because of independent assimilation processes, are possible according to the constraints on syllabic codas: in fact, *k* and *l* can be found in codas in sequences such as *kt* or *lt*. Therefore, epenthesis does not occur to remove consonantal sequences that do not occur in the language, but just to remove violations of syllable structure constraints. It follows that the epenthesis patterns can only be explained if the L1 syllable structure constraints, rather than the impossibility of the consonantal sequence, influence the perception of consonant clusters. It also follows that these patterns are due to structural (featural) differences, rather than frequency differences (e.g., Vitevitch & Luce, 1998). If perceptual epenthesis were a means by which the perceptual system adjusts clusters that presumably have zero frequency, then all the illicit consonant clusters in Korean would be more susceptible to epenthesis, which is contrary to what one finds (See Moreton, 2002 for similar results in the case of English listeners.)

These perceptual illusions appear to have a clear grammatical, phonological motivation. In the literature they are usually referred to as 'phonetic illusions' (Dupoux et al., 1999; Dupoux, Pallier, et al.,

2001; Dupoux, Peperkamp, et al., 2001), but I believe that they would be better described as 'phonological illusions', and this is the term that I will use in this paper.

Phonological illusions are obviously not restricted to the perception of foreign syllabic structures. In fact, it is known at least since Polivanov (1931) that listeners have problems with non-native, foreign sounds and misperceive them, i.e., have an illusory perception of their reality, and a number of experimental studies have also shown that (Best, 1994, 1995; Best, McRoberts, Lafleur, & Silverisenstadt, 1995; Werker & Lalonde, 1989; Werker & Logan, 1985; Werker & Tees, 1984).

Evidence for perceptual distortion of non-native phonological configurations is also provided by nativization processes in loanword phonology (Boersma & Hamann, 2009; Calabrese, 2009; Peperkamp & Dupoux, 2003; Peperkamp, Vendelin, & Nakamura, 2008; Silverman, 1992). Linguists have in fact observed that there is a strong parallelism between the attested processes of nativization found in a language and the perceptual distortions implemented in experimental settings by subjects speaking that language. For instance, Japanese speakers are known to insert epenthetic vowels in consonantal clusters not allowed in Japanese when they pronounce loanwords, or more generally foreign words ([*ma.kɯ.do.na.rɯ.do*] 'Mac Donald', [*sɯ.to.ra.i.ko*] 'strike'), and in the same way Japanese listeners perceive illusory epenthetic vowels in such clusters, as discussed above. The same phenomenon occurs in the perception/production of non-native sound contrasts; for example, Korean listeners find it hard to distinguish between the English consonants [r] and [l] in CV-stimuli (Ingram & See-Gyoon, 1998), and in loanwords from English, word-initial [l] is adapted as [r] (Kenstowicz & Sohn, 2001). In a similar vein, French listeners have severe difficulties with perceiving stress contrasts (Dupoux, Pallier, Sebastián-Gallés, & Mehler, 1997) and in loanwords, stress is systematically word-final, regardless of the position of stress in the source word.

There is thus widespread agreement among linguists that nativization processes occur in perception (see LaCharité & Paradis, 2005 for a dissenting view). Furthermore they involve clear phonological operations: the misperception cannot be accounted for in terms of phonetic or acustic similarity but in terms of phonological operations on abstract configurations built in terms of distinctive features. For example, the nativization of English lax high vowels /I/ and /U/ into French /i/ and /u/ instead of the acoustically closer /e/ and /o/ (LaCharité & Paradis, 2005) can be readily accounted for in terms of operations on distinctive features where the configuration [+high, −ATR] of English /I/ and /U/ is repaired into the configuration [+high, +ATR] of the French high vowels /i/ and /u/ (see Calabrese & Wetzels, 2009; Kang, 2010 for a review of other cases of this type). Examples of "phonological" nativizations abound. The Italian vowel system does not have the [+low, −back] vowel /æ/ of the English word /kæt/ and Italians interpret this vowel as either [e] or [a] by replacing [+low] with [−low] or [−back] with [+back].

The same obviously holds for consonants. For example, the retroflex stops /ʈ, ɖ/ of Hindi, featurally [+distributed, −anterior] stops, are usually misperceived as anterior coronal stops [t, d] by Italian or English listeners. This misperception is accounted for by replacing the specification [−anterior] with [+anterior]. At the same time, they may also be misperceived as postalveoalar stops (which are affricated) [tʃ, dʒ] by replacing [−distributed] with [+distributed].

More complex phonological changes can also occur. Kim (2009) reports that the labial fricative /f/, featurally [+continuant, +labial], may be misperceived as either /p^h/ or /hw/. The first misperception can be obtained by changing the specification [+continuant] into [−continuant] and in addition inserting the feature [+spread glottis] (aspiration)—because English voiceless stops are interpreted as aspirated in Korean. The misperception /hw/ for /f/ would be an instance of the phonological operation of fission (Calabrese, 2005) which takes a feature combination realized in a single sound and distributes it into two different sounds: [+continuant] in h and [+labial] in w. The same thing occurs in the case of English [+continuant, −anterior] [ʃ], which may be misperceived as the sequence [s + j], i.e., ([+continuant]) + ([−anterior]) in Korean (Kim, 2009). In the same vein, notice that English listeners misperceive Italian palatal nasal ([+nasal, −anterior]) [ɲ] as the sequence [n + j], i.e., ([+nasal])+([−anterior]), so that lasagna [lasaɲna] is misperceived as [lasanja].

All of these misperceptions demonstrate that the perceptual process involves access to abstract phonological computations. Perception is not direct, but mediated by grammatical/linguistic knowledge. It is proposed below that this access to phonological computation can be readily accounted for if

perception involves an analysis-by-synthesis procedure, which requires an active access to grammatical knowledge and elaboration of perceptual targets through grammatical derivations.

3. Perception

A sentence, when uttered, is only a stream of sound. That stream of sound, however, has associated with it a certain meaning.

In producing an utterance, a speaker converts a determined conceptual structure into a stream of sounds. In perceiving an utterance, a listener converts a stream of sounds into a conceptual structure. The speakers' and listeners' linguistic knowledge of a given language must contain information that is able to account for how sound and meaning for the sentences of his language are correlated, or for how conversions between sound and meaning occur.

It is commonly assumed that knowledge of words, or more precisely of the vocabulary of the language, is a fundamental part of this knowledge. Words, on their turn, are commonly composed of smaller pieces, morphemes. And it is the morphemes—in addition to the words—that make up the vocabulary of the speakers of a language. Each vocabulary item is composed of a phonetic index, a sequence of phonemes, what I will call here an exponent, and an associated meaning, a conceptual unit or a combination of conceptual units. It is self-evident that the knowledge of exponents must be stored in the memory of speakers: we are not born with this knowledge, but must learn them; i.e., commit them to memory one by one.

By means of exponents, hierarchically organized structures composed of grammatical and semantic features generated by a syntactic computation—what we can loosely call conceptual structures—are converted into phonological representations. This is done by associating the exponents to the relevant morphological pieces in these structures—a process referred to as the insertion of exponents. The phonological and phonetic components then convert the phonological representations generated by the insertion of exponents into articulatory representations that are then implemented in patterns of muscular activations/articulatory gestures. At this point streams of sounds are produced.

The objective of listeners is to access the meaning conveyed by the stream of sounds occurring in the utterances they hear. The meaning of an utterance is accessed through the identification of the exponents of the vocabulary items (morphemes/words) used in it and the recognition of how they are structurally organized in this utterance.[2]

A crucial assumption is that the representations of exponents of the morphemes and words stored in long-term memory involve phonological distinctive features (see Section 5). Therefore, to achieve identification of the exponents present in the input signal, listeners must convert the signal into an abstract phonological representation structured in such a way that it can interface with the feature-based representations of the exponents stored in long-term memory. These abstract representations are most plausibly constructed through a multi-time resolution processing of auditory representations, which generates preliminary, broadbrush hypotheses about the featural content of the representations (Boemio, Fromm, Braun, & Poeppel, 2005; Poeppel, 2001, 2003; Poeppel, Idsardi, & van Wassenhove, 2008; see below for more discussion).

I assume that once the abstract representation of the input is constructed, access to meaning is achieved in the following way. The identification of the exponents of the morphemes and words present in the input signal is done by matching the hypothetical exponents extracted from the abstract representation of the input with the exponents of morphemes and words that are stored in a long-term memory vocabulary (see Marslen-Wilson, 1987; McClelland & Elman, 1986 on how small chunks of input can be used to activate stored lexical representations). This matching process is then followed by a top-down construction of a sentence that tries to make sense of the morphemes and words that have been identified in the input. When a vocabulary item is chosen, its morpho-syntactic features and its

[2] The term 'speech perception' is used in the literature to refer only to the processes culminating with the identification of the exponents of words and morphemes whereas the term 'language comprehension' is used to refer to the general process by which streams of sounds are converted into meaning. For the sake of simplicity, in this paper I will use the term speech perception also to refer to what is otherwise called language comprehension as in the preceding paragraph.

meaning are accessed in the vocabulary and in the encyclopedia (a more general container of stored linguistic knowledge; see Section 7). These morpho-syntactic features and these meanings are checked in the context of the morpho-syntactic and semantic structures that are in the meanwhile generated to account for the order of the elements and the phrasing in the representation of the utterance. The meaning of the generated structures is also checked against the general pragmatic context. I assume that all of these processes/derivations occur in parallel and will eventually converge in producing a morpho-syntactically and semantically well-formed sentential representation, the "meaning" of the utterance.

Observe that the exponents are only a medium to access or convey the meaning. They are like ladders that, once used for their purpose, can be left behind. Once the syntactic and semantic representations of the sentence are constructed, they can be disregarded. The phonological composition of these exponents is not important other than for its function of making the meaning accessible. In fact, when we hear an utterance, we do not remember the exact words that a speaker used, but rather the meaning we gleaned from them (Sachs, 1967).

In certain situations, however, the listener may want to be certain that the exponents identified in the input signal correspond to those intended by the speaker who produced the signal. The need for this certainty is referred to as the parity requirement (Liberman, 1996; Liberman & Whalen, 2000). In conditions not requiring parity, i.e., when the listener is dealing with well-known, familiar configurations in the input signal, the broadbrush hypotheses about the featural content of the representations mentioned earlier may provide enough information for a certain identification of the exponents present in the input signal. No additional attention needs to be paid to the exponent, and the meaning behind it can be simply accessed, as discussed above.

In situations requiring parity, however, the goal of speech perception is the exponent more than the meaning that is behind it; not only does the complete featural composition of the exponent need to be ascertained, but also its articulatory implementation. As discussed in sect. 5, features are abstract correlations between auditory and articulatory properties and must eventually be converted into concrete articulatory representations.

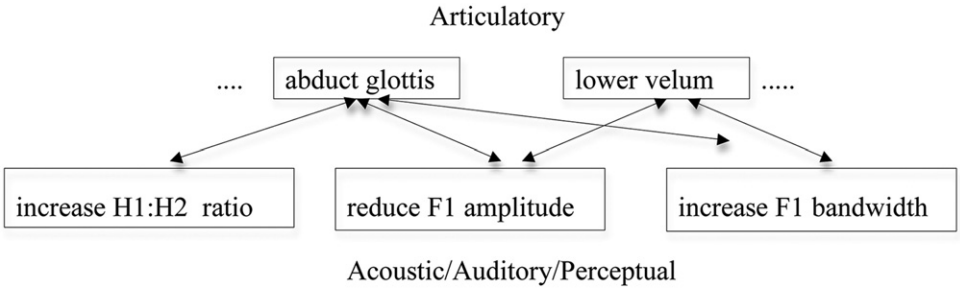
Parity requirements are needed for example when a novel signal is met—in the case of familiar, known signals one can assume that the listener is certain of what exponents are in the input. This obviously occurs when we need to learn a new word in L1 or L2. In this case, the identification of the phonological structure of the exponent is the main goal of the perceptual process so that we can properly memorize it. In this case the input must be analyzed and converted into an articulatory representation. A mental representation of the phonological structure of the exponent in terms of articulatory features must be constructed. Achieving parity requires testing whether or not this hypothetical representation matches the input signal. Parity requirements are also needed when the listener hears a novel utterance where known words and morphemes may be combined in novel ways, and thus may have a novel shape due to phonological processes applying across words and morphemes, or when familiar constructions or sentences are pronounced by a non-native speaker or a speaker with a different accent. Observe that in this case the hypothetical representation that is constructed includes hypotheses on the exponents of words and morphemes underlying the utterance (see below for discussion of an example). Parity requirements must also be met in experimental conditions when the subjects are performing tasks requiring the identification of non-lexical or sub-lexical aspects of the stimulus. Essentially, we can say that all of the situations requiring parity involve sublexical identification of the phonological structure of the input string present in the signal. This is true also for cases where exponents are present insofar as only their phonological structure matters in this case, and not the access to their meaning.

Approaches such as Direct Realism (Fowler, 1986) assume that parity can be achieved directly. According to these approaches, in fact, articulatory representation can be directly extracted from the acoustic signal through a direct recovery—bottom-up—of the articulatory configurations behind the acoustic signal.

This proposal faces problems. First of all, as discussed by Diehl, Lotto, and Holt (2004), there is the fact that articulatory configurations cannot be recovered uniquely from the acoustic signal. In fact, there are many different ways to produce the same speech signal. For example, approximately the same formant pattern can be achieved either by rounding the lips, lowering the larynx, or doing a little of both (Riordan, 1977). Evidence that different articulatory gestures can generate similar acoustic

patterns is presented in Ohala (1996), Ladefoged, DeClerk, Lindau, and Papcun (1972), and Nearey (1980). In addition, the same articulatory gesture can have different acoustic effects. For example, Halle and Stevens (1971) drew attention to the fact that differences in vocal cord stiffness have vastly different effects in sounds produced with a small pressure drop across the vocal cords (e.g., obstruents) than in sounds produced with a large pressure drop across the vocal cords (e.g., vowels). Therefore, as has long been recognized, there is a complicated, many-to-many relationship between motor speech actions and their acoustic and auditory consequences; a small fragment of this mapping is shown in (1) (after Idsardi, 2007).

(1)



As discussed by Diehl et al. (2004), the problem of the extraction of articulatory representations from the acoustic signal can be solved only by assuming access to the knowledge of the speech production apparatus, top-down, as in the Motor Theory of Speech Perception (Liberman & Mattingly, 1985, 1989). Access to this knowledge is also an aspect of the analysis-by-synthesis model discussed below.

On the other hand, phonological illusions such as those discussed in Section 2 also show that the construction of the articulatory representation underlying the speech signal must be indirect and mediated—top-down—by previous (linguistic) knowledge. Facts such as those we discussed about Japanese and Korean show that the identification of sounds in acoustic inputs needs information that is not contained in these inputs, and that therefore cannot be simply extracted bottom-up from them, rather must be computed—top-down—through processes that can access the grammatical system.

Interestingly, "illusions" or misperceptions are also observed in the experience of the native language. Kenstowicz (1994), for example, reports that English speakers tend to perceive the inter-syllabic consonantal material in *camper* and *anchor* as analogous to *clamber* and *anger.* This is an illusion, however. In most dialects (Malecot, 1960) the nasal consonant is phonetically absent before such sounds as [p,t,k,s], so that *camper* and *anchor* have the same gross phonetic shape (C)$\tilde{V}$CVC ($\tilde{V}$ = a nasal vowel) as (C)VCVC *wrapper* and *acre.* While $\tilde{V}$CVC *anchor* belongs with VCVC *acre* phonetically, English speakers have the strong intuition that psychologically it belongs with VCCVC *anger.* As Kenstowicz observes, this perceptual judgment is based on the abstract phonological representations of these words.

Access to top-down information in the construction of perceptual representation of the signal is shown by other facts. In fact, we know that listeners may "restore" missing phonetic segments in words (Samuel, 1981; Warren, 1970), and talkers shadowing someone else's speech may "fluently restore" mispronounced words to their correct forms (e.g., Marslen-Wilson & Welsh, 1978). This ability to restore missing phonemes or correct erroneous ones can be explained only if we assume top-down processes that access information in the lexical entries.

Even grosser departures of perceptual experience from the stimulus may be observed in some mishearings (for example "popping really slow" heard as "prodigal son" (Browman, 1980; Fowler, 1986) or "mow his own lawn" heard as "blow his own horn" (Fowler, 1986; Garnes & Bond, 1980)). As for mishearings, Garnes and Bond (1980) argue that "active hypothesizing on the part of the listener

concerning the intended message is certainly part of the speech perception process. No other explanation is possible for misperceptions which quite radically restructure the message …" (p. 238).

In conclusion, access to grammar, access to lexical information, and the ability to reconstruct, or even construct, phrases and sentences in the percept, in addition to grammatically induced phonological "illusions", indicate that perception clearly involves top-down processes, and that articulatory representations cannot be accessed directly as proposed by Direct Realism.

As recent work on speech perception has suggested (Bever & Poeppel, 2010; Poeppel & Idsardi, 2010; Poeppel et al., 2008; Poeppel & Monahan, 2010) (see also Calabrese, 2009), the most adequate way to account for the aforementioned "active hypothesizing on the part of the listener", for the effects of grammatical computations, and in general, for the interaction of top-down and bottom-up processes in speech perception, is to assume that perceptual representations of exponents are constructed through an analysis-by-synthesis of the signal (Halle & Stevens, 1962). According to this analysis-by-synthesis model, the listener analyzes the acoustic input by deriving how it is generated by the speaker, synthesizes a virtual acoustic signal based on the output of this derivation, and matches the virtual to the actual signal. The first step is the generation of a hypothetical representation underlying the perceptual target representation. The hypothetical representation is then submitted to a computation that derives a representation that can be compared with the target.

Let us consider a case where parity is required: the perception of connected speech. Just consider the following sentence as pronounced in American English:

(2)

a) How did you find it and hit it after I hid it?
b) [hawdɪjəfɑ̃ɪDIDɔ̃dhIDIDaftəraɪhIːDi?]

The phonetic transcription in (2b) demonstrates some of the problems involved in the perception of this utterance insofar as there cannot be any 'direct' mapping from the speech input to the exponents of lexical entries not mediated by the application of phonological rules. For example, the palatalization of final /d/ before /y/ in /did/ means that any attempt to relate that portion of the speech input to the exponent /did/ is likely to fail. Similar points can be made about the flapping and glottalization of the phonemes in /hit/ and /it/. Furthermore in the case of the strings [hɪdɪd] and [hɪːdɪdi?], two phonological processes, vowel lengthening before voiced stops and flapping, interact in such a way that the crucial phonemic contrast between the /t/ of /hit/ and the /d/ of /hid/ is reduced to a phonetic difference in vowel length thus making any direct access to the relevant lexical items impossible. In addition, (2) illustrates the well-known point that there are no reliable phonetic or phonological cues to word boundaries in connected speech. Without further phonological and lexical analysis there is no indication in a transcription like (2b) of where words begin or end; for example, how does the lexical access system distinguish word-initial /I/ in /it/ from word-internal /I/ in /hid/? The most adequate way to analyze and reconstruct the lexical/morphological composition of the utterance is the following: Given a rough phonological analysis of the input, hypotheses are made on the words and morphemes present in the input, specifically on their exponents. The generation of these hypotheses is aided by the knowledge of the phonological processes present in the language, which allows parsing out the effects of phonological processes. Once these hypothetical underlying representations are established by activating long-term memory representations of morphemes and roots, phonological processes are applied to them in the relevant order to generate a hypothetical surface representation. If this hypothetical surface representation matches the actual surface representation of the utterance and its auditory source, parity is satisfied and identification and recognition can safely occur. If there is no match, hypothesis generation and subsequent grammatical construction of the surface representation must be attempted again. One can assume that all hypothesis generations and subsequent derivations run in parallel until a successful match is reached.

The idea that speech is perceived by accessing phonological computations assumes that in the perception process, the listener has an active role: he is able to access abstract morphological and syntactic levels of representation of the perceived utterance and compute the surface articulatory

shape of the utterance from these abstract levels. A successful perceptual act occurs when the acoustic shape of the articulatory representation derived in this perceptual computation matches the acoustic input in auditory memory.[3,4]

During the analytic process characterizing analysis-by-synthesis, any type of knowledge can be accessed. In particular, the listener's ability to access the knowledge of the speech production apparatus can explain how the articulatory configurations behind the signal are properly identified, as proposed by the motor theory of speech perception (Liberman & Mattingly, 1985, 1989; see also Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). In the same way, the ability to access knowledge of the speech production mechanisms accounts for the existence of illusions like the McGurk effect (van Wassenhove, Grant, & Poeppel, 2005). Access to lexical knowledge during this analytical process, on the other hand, explains phoneme restorations and mishearings. Observe that under this analytical mode, perceptual representations are constructed indirectly and on a hypothetical basis. Therefore, they can be mistaken and may actually not reflect what is in the input. This explains the possibility of illusions.

I assume that the retrieval of the exponent's phonological representation from the acoustic signal proceeds as follows. Mesgarani, David, Fritz, and Shamma (2008) show that the spectro-temporal receptive fields in core auditory cortex permit the construction of a relatively high fidelity representation of the signal. These auditory representations contain a rough spectral analysis that is further analyzed in subsequent, more abstract representations. Poeppel et al. (2008) propose that the mapping from a spectro-temporal acoustic representation to a lexical-phonological one is mediated by an intermediate representation named *phonological primal sketch* (PPS). This representation is obtained by 'sampling' the output of core auditory cortex using two temporal window sizes. One window is on the order of 20–50 ms and highlights the rapid temporal changes occurring in the signal, another is on the order of 200 ms and represents the narrow band frequency variation (formant structure). (Boemio et al., 2005; Poeppel, 2001, 2003) This two-stage sampling of the auditory spectral representation corresponds to the two-stage sampling proposed by Stevens (2002), which I discuss here. According to Stevens, first the locations and types of basic acoustic landmarks in these representations are established. These acoustic landmarks are identified by the locations of low-frequency energy peaks, energy minima with no acoustic discontinuities, and particular types of abrupt acoustic events. From these acoustic landmarks certain articulatory events can be hypothesized: the speaker produced a maximum opening in the oral cavity, or a minimum opening without an abrupt acoustic discontinuity, or a narrowing in the oral cavity sufficient to create several types of acoustic discontinuity (Stevens, 2002). Such landmarks provide evidence for distinctive features such as [consonant], [sonorant], [continuant], [strident], the so-called articulator-free features (Halle, 1995). In this way, the basic syllabic structure of the sequence is identified. The second step consists of the extraction of acoustic cues from the signal in the vicinity of the landmarks. These cues are derived by first measuring the time course of certain acoustic parameters such as the frequencies of spectral peaks or spectrum amplitudes in particular frequency ranges, and then specifying particular attributes of these parameter tracks (Stevens, 2002). These acoustic parameters provide evidence for which articulators are involved in producing the landmarks, and how these articulators are positioned and shaped. Stevens (2002) proposes that once the sequence of landmarks has been identified, and a set of acoustic cues has been evaluated in the vicinity of each landmark, the next step is to convert this landmark/cue pattern into a symbolic or quantal description consisting of a sequence of feature bundles. Stevens hypothesized that this conversion is carried out by examining acoustic landmarks and adjacent cues and analyzing them in term of their overall environment and prosodic context. At this point guesses concerning the feature specifications (see Stevens, 2002 for more discussion) must be generated. In this way, the auditory representation of an utterance is converted into a hypothetical array of features. One can hypothesize that in conditions not requiring parity, i.e., when the listeners is dealing with well-known, familiar

[3] As Poeppel et al. (2008) observe, analysis-by-synthesis should be conceptually related both to internal forward models that use predictive coding and to Bayesian classification approaches.

[4] The 'forward' synthesis of candidate representations makes the processing characterizing this model completely active, and therefore intrinsically different from the "passive" processing typical of bottom-up models where analytic features percolate up the processing hierarchy or of connectionist-style models where activation patterns simply spread (Poeppel et al., 2008).

configurations in the input signal, this hypothetical array of features provides enough information for a certain identification of the exponents present in the input signal.

However, as discussed above, when parity is needed, and the correctness of the hypothesized representations extracted from the auditory representation of the utterance must be checked, the hypothetical array of features must be submitted to a phonological computation that generates a surface representation. This representation must then be converted into a format that permits comparison with the input signal, plausibly stored as an auditory representation. Therefore it must be converted into an auditory representation itself by computing the auditory correlates of the acoustic effects of its articulatory implementation. If this auditory representation matches the auditory representation generated from the acoustic signal, the mental representation is accepted as the correct perceptual representation.

4. Two modes of speech perception

It is to be noted that the objective of listeners is to access the meaning of the utterance they hear, and that as argued before, the exponents are only a means to access or convey this meaning. Once the syntactic and semantic representations are constructed, they can be disregarded. Therefore the best way to achieve this objective is to do it fast and with minimal expenditure of resources. This may be obtained by searching auditory representations only for auditory patterns and properties corresponding to the idiosyncratic, contrastive properties of phonological representations, i.e., for all those formal properties of exponents that can be associated with differences of meaning in the language: idiosyncratic collocation of segments in different syllabic structures, contrastive features in segments, contrastive differences in prosodic structures, tonal contrasts. In this case attentional resources in perception are restricted only to those patterns of the auditory representation that correlate to these contrastive phonological properties. This is what Werker and Logan (1985) call phonemic perception. One can hypothesize that the construction of the phonological sketch involves only phonemic perception, and that the typical mode of perception is phonemic (see Sapir, 1933). In such a mode, noncontrastive aspects of perceptual representations are neglected. Indeed, recent studies (Eulitz & Lahiri, 2004; Hwang, Mohanan & Idsardi, 2010; Kazanina, Phillips, & Idsardi, 2006; Lahiri & Reetz, 2002, 2010) show that normally perceivers are not sensitive to, or do not access, noncontrastive aspects of perceptual representations.

But if perception were only phonemic, all allophonic variation would be imperceptible, inaccessible and therefore unlearnable. If allophonic variation were always due to universal adjustments due to coarticulation, this would not be a problem. However, there is also allophonic variation that is language-specific, like the aspiration of voiceless stops in foot initial position in English. In order for English speakers—and also non-native ones—to learn that such a process is present in this language, they must be able to access it. This would be impossible if perception were only phonemic. Therefore, noncontrastive details of phonological representations must be accessible. Now, such details are indeed available in the high fidelity representations of the signal constructed in the spectro-temporal receptive fields of the core auditory cortex mentioned before. It is plausible to assume that they can be accessed there if attention is paid to them. This is what Werker and Logan (1985) call phonetic perception.

According to Werker and Logan, when subjects perceive stimuli according to native-language phonological categories, they are demonstrating "phonemic" perception. When subjects instead show a sensitivity to phonetic distinctions, they are using phonetically relevant (or "phonetic") perception.[5] This perceptual mode is necessary not only to learn allophonic variation in the native language, but also to access sound contrasts present in other languages, and to learn foreign sounds and sound configurations. Furthermore, it is also necessary in all of the cases where parity requirements must be met. In these cases, in fact, the perceiver must be able to access all aspects of the input representation in order to be certain that the representation that he is constructing mentally is the correct one. We will come back to these two perceptual modes in Section 8.

[5] As a phonology teacher in a US academic institution, I am always surprised by the fact that English students are not aware that they aspirate voiceless stops in foot initial position. However, this lack of awareness is simply explained by the fact that humans' normal mode of perception is phonemic. Sure enough, once the students are told that their voiceless stops are aspirated in that position, and pay attention to that property—i.e., once they are instructed to use a phonetic mode of perception—they readily discover this characteristic trait of their speech.

It is to be noted in this regard that the animal auditory system must satisfy two conflicting needs (Nelken, 2008). The first need is that of encoding sounds in all their details. A highly detailed representation of sounds is necessary for detecting informative, but physically small and context-dependent, changes in the incoming sound stream. However, much of this detail is irrelevant in any given situation. The second need therefore is that of extracting the relevant information out of the highly detailed representation of the sound stream in order to encode sounds in terms that are useful for guiding future behavior. The relevant information may be very different for different tasks. In the case of human speech perception these two needs may be associated with the two perceptual modes mentioned above: phonetic perception is associated with the first need, phonemic perception with the second one.

5. Features

Being a linguist, up until now I have assumed that sounds are bundles of distinctive features. In this section, I briefly discuss some evidence for this assumption. I will also consider some aspects of the ontology of distinctive features.

The majority of linguists assume that speech sounds are not the ultimate atoms of linguistic organization but that they can be decomposed into more fundamental categories grounded in the structure and behaviors of the vocal apparatus and of the sensory system. Any particular sound is a complex of these categories, which linguists call distinctive features.

The role of features in phonology was first recognized in the earliest works in phonology when it was observed that sound systems are structured in terms of regular correlations based on recurrent elementary components, which were later called distinctive features (e.g. Jakobson, Fant, & Halle, 1952; Trubetzkoy, 1939). In a recent article (2009), Clements provides new and compelling evidence for assuming that indeed the structuring of sound systems can be adequately accounted for only in theories hypothesizing that sounds are bundles of features taken from a universal set of phonetic properties that are exploited to create contrasts.

Work mostly in generative phonology has elaborated the nature and role of features in many additional areas of phonology, especially as the essential elements that account for phonological patterning in the distribution of sounds and in phonological processes (see Halle, 2002; Kenstowicz, 1994; Mielke, 2008; Vaux, 2010; a.o., for more discussion of this aspect of feature theory). A striking finding about this patterning is that while certain groupings of sounds are found recurrently in phonological processes, others are never attested. Now it so happens that the recurrent groupings of sounds—called natural classes of sounds—can be specified as containing a small set of features, whereas those that are not recurrent cannot be specified in any simple way. If the sounds in phonological representations were like the symbols of the IPA, unitary and unanalyzable elements, we would not be able to account for this fact. However, if sounds are represented as complexes of features, certain sets will be relatively simple to specify; {*u, o, a*} = [vowel, +back], {*i, e, æ*} = [vowel, −back], {*b, d, g*} = [stop, voiced][, {*p, t, k*} = [stop, voiceless], {*m, n, ŋ*} = [consonant, nasal]. On the other hand, rules characteristically relate sounds as inputs and outputs in a nonrandom way. There are rules like the following: [*u, o, a*] → [*i, e, æ*]. But there are no rules like [*u, o, a*] → [*æ , i, e*] or, worse, [*a*] → [*s*] because such rules cannot be accounted for in terms of simple featural changes.[6]

In order to appreciate the role of features in phonology, it is important to observe that in addition to processes in which a segment assimilates a feature of an adjacent segment, there are also processes that operate just on features without a segmental trigger. A process of this type (Calabrese, 2010) is found in many Italian dialects where for example the marking of the 2nd person singular is obtained solely through the application of the spreading of the feature [+high] to the stressed vowel of the verbal stem (cf. 'vedə/'vidə "see-1sg/2sg., 'korrə/'kurrə "run-1sg/2sg.", 'mettə/mittə 'put-1sg/2sg.'" in the dialect of Arpino in southern Italy). Note that processes of this type are not restricted to word-

[6] Not all processes across languages can be accounted for in terms of natural classes. There are indeed processes involving unnatural classes, the so-called crazy rules (Bach & Harms, 1972). It can be shown, however, that these processes are the fossilized outcome of processes that originally involved natural classes but that were later modified due to the cumulative effect of changes normally affecting the history of languages (see Mielke, 2008 for recent discussion of the role of features in relation to crazy rules).

internal environments only but also occur when words are sequenced in syntax. For example, in Chaha (McCarthy, 1983), the third-person singular object is indicated with labialization on the verb: *dænæg/dænæg*^{w} 'hit-*with object/with no object*', *mækær/mæk*^{w}*ær* 'burn-*with object/with no object*'. In Welsh 'direct object mutation' (Borsley, 1997, 1999) the initial stop of a direct object of a finite verb becomes fricative, i.e., it changes from [−continuant] to [+continuant], while there is no such change when the verb is nonfinite (cf. ***f****eic* vs. ***b****eic* 'bike'). Now, there is no way that previous memorization and listing of words can account for the generation of these forms, insofar as novel syntactic constructions can always be built so that a listener may have never encountered the affected words before. The alternations we observe in these cases can be accounted for only by assuming that they are due to phonological operations that change the featural composition of the morphemes and words involved in these constructions. Therefore, the hypothesis that morphemes are constructed in terms of distinctive features, and may even consist of a single feature, is fundamental here.

From what we have discussed up to now, we can conclude that segments in mental representations are complexes of features. Features are not a convenient classificatory system for sets of whole-sound fundamental speech units (phones or phonemes) as implied by the IPA chart, but they are the fundamental units themselves. The issue that must be addressed now is the nature of these features. Since Chomsky and Halle (1968), linguists overwhelmingly assume that features have an articulatory basis. Phonological analysis of language after language shows that classes of sounds appear to be organized in terms of the articulatory correlates of features (see Halle, 2002; Halle, Vaux, & Wolfe, 2000 for recent arguments).

The role of articulatory features in production is obvious. The issue is how representations based on such articulatory features are recovered in perception. As already mentioned, these representations cannot be directly extracted from the acoustic signal since articulatory configurations cannot be recovered uniquely from the acoustic signal. As discussed before, this problem can be solved only by assuming that the listener has access to knowledge of the basic correlations between motor speech actions and acoustic and auditory patterns, top-down, through an analysis-by-synthesis model of speech perception. Therefore, analysis-by-synthesis is necessary not only to account for the relations between hypothesized featural representations of exponents and what is contained in the signal, but also for the relations between features and the acoustic information contained in the signal (see Hawkins, 2009 for interesting discussion of this point).

A concept important for understanding how articulatory patterns behind these auditory properties/patterns are identified is the notion of quantal property of sound. Stevens (1972, 1989) has shown that the relations between the articulatory and the acoustic representations of speech have certain quasi-discrete or quantal characteristics that are exploited in speech production and perception. In particular, acoustic studies (Stevens, 1972, 1989, 1998) of sounds produced by various manipulations of the vocal tract show that certain stable acoustic patterns occur when the vocal tract is in particular configurations or performs particular maneuvers—these configurations or maneuvers correspond to distinctive features. As Stevens (2002) points out, these combinations of acoustic and articulatory patterns are based on the physics of sound generation in the vocal tract, including theories of coupled resonators, the influence of vocal-tract walls on sound generation, and discontinuities or stabilities in the behavior of sound sources. Quantal relations establish correlations between acoustic patterns and articulatory patterns. The features are nothing else than these correlations. For example, the feature [+round] would define the connection between the motor gesture of lip-rounding (i.e., the enervation of the orbicularis oris) and a particular perceptual pattern, perhaps the down-sweep in frequencies across the whole spectral range. Similar correlations between articulatory activity and acoustic signal are provided for each of the nineteen or so features that make up the universal set of phonetic features (cf. Halle, 2002; Halle & Stevens, 1991). These quantal attributes help to simplify the process of uncovering the discretely specified segmental and categorical representations that are the building blocks of words (Stevens, 2002).

Features also have a grounding in auditory representations. The ability for the listener to analyze the acoustic signal into discrete categories in auditory representations—an ability that is fundamental for featural distinction in perception—is demonstrated by several studies that suggest that complex events such as stream segregation—extracting the abstract sound patterns and invariant sound relationships—and categorical speech perception may take place in the auditory cortex (Dehaene-Lambertz, 1997; Näätänen, 2001; Paavilainen, Jaramillo, Näätänen, & Winkler, 1999, 2001; Phillips, 2001;

Phillips et al., 2000; Shestakova et al., 2002; Sussman, Fruchter, Hilbert, & Sirosh, 1999, Sussman, Winkler, Huotilainen, Ritter, & Näätänen, 2002; Tervaniemi, Saarinen, Paavilainen, Danilova, & Näätänen, 1994). The same thing is shown by quantal aspects of auditory responses to sound, such as responses to acoustic discontinuities and to closely spaced spectral prominences (Chistovich & Lublinskaya, 1979; Delgutte & Kiang, 1984; Stevens, 2002).

Animal research shows that the capacity to analyze the acoustic signal into discrete categories in auditory representations is already found in animals. Mesgarani et al. (2008) show that neuronal responses to continuous speech in the primary auditory cortex of animals (ferrets, in their study) reveal an explicit multidimensional representation that is sufficiently rich to support the discrimination of human phonemes. This representation is made possible by the wide range of spectro-temporal tuning in the primary auditory cortex to stimulus frequency, scale, and rate. The great advantage of such diversity is that there is always a unique subpopulation of neurons that responds well to the distinctive acoustic features of a given phoneme and hence can fully analyze that phoneme in its featural complexity. One can therefore speculate that the human ability to use features in perception can in principle be explained in terms of basic properties of the auditory representations common to many mammalian, and also avian, species (Kuhl, 1991). Some investigators have argued that the innate representations that infants bring to the language-learning task simply reflect the natural boundaries found in the sophisticated auditory representations that humans share with other species (Kuhl, 1991).

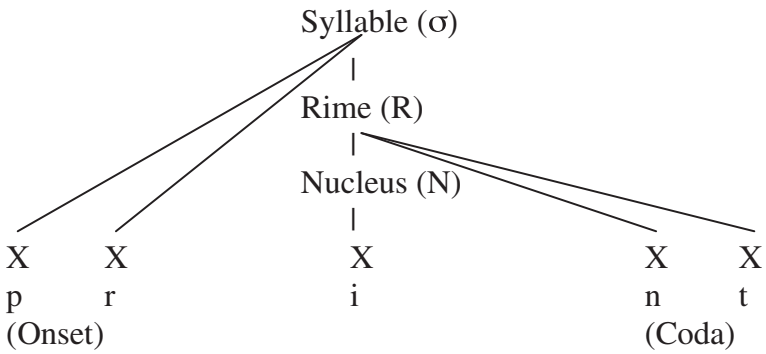
6. Syllable structure

To understand the phonological illusions we observe in Japanese speakers it is important to briefly discuss syllable structure.

Segments are grouped into syllables so that words are sequences of syllables, see dis+or+ga+ni+za+tion, an+ti+pre+des+ti+na+ri+a+nism, hair+breadths. Syllables are the locus of the sequential constraints on concatenation of sounds. Such sequential restrictions are among the most striking differences among languages. For example, English allows a great variety of consonant clusters both in prevocalic position and in postvocalic position of the syllable, whereas Japanese requires that there be precisely one consonant in the prevocalic position and one in postvocalic position with the restriction that this consonant must be the first member of a geminate or a nasal consonant.

The hierarchical syllabic structure I will adopt here is given in (3) for the word *print* (Blevins, 1995). The Xs are called skeletal positions and represent abstract syllabic positions; the letters /p, r, i, n, t/ stand for complexes of features[7]:

(3)



[7] The initial or final /s/ found before or after a stop in *sprint, sprints* or *stop, stops* has a special status in syllable structure. I will not discuss it here.

Every syllable must have a nucleus (N). According to the language, the other syllable constituents may be optional. The nucleus position is typically filled in by vowels; in some languages such as English, sonorant consonants can also appear there (*button* [[bʌtn̩]], *kettle*[[ketl̩]]). All languages permit an onset (O); a fair number of languages require it. Many languages permit a coda (C); a number of languages prohibit the coda and require that all syllables end with a nucleus (i.e., a vowel). It follows that the basic syllable present in all languages has an onset, a nucleus and no coda, i.e., is a CV syllable, where V represents a nucleus and C a syllable margin (an onset or coda depending on the position with respect to the nucleus). Some languages allow onsetless syllables, i.e., CV, V. Others allow codas: CV, CVC, or CV, V, CVC, VC, if they also allow onsetless syllables. Finally, some languages additionally allow complex onsets, i.e., onsets composed of two or more segments, e.g., CCV, and others also allow complex codas. This is the case of English, which has the following simplified syllable inventory: CV (me), CCV (*tree*), V (a), CVC (*tip*), CCVC (*trip*), CVCC (*kart*), CCVCC (*print*), VC (*am*), VCC (*art*). Italian, on the other hand, does not allow complex codas, but has trisegmental onsets (*pro. prye.ta*).

Since Clements and Keyser (1983) it has been hypothesized that the absence of a given syllabic configuration in a language is accounted for in terms of a negative constraint. For example, if a language does not have a complex onset, it is hypothesized that this is the result of the activity of the constraint *CCV; if a language does not have a complex coda, this would be the result of the activity of the constraint *VCC. In contemporary models of phonology starting from Calabrese (1988), Paradis (1988) and more recently in Optimality Theory (Prince & Smolensky, 1993/2004), a fundamental role in phonological computations/derivation is played by constraints such as *CCV and *VCC. In Calabrese (1995, 2005) constraints such as these are called marking statements. They identify configurations that are phonologically difficult or complex and that may be found in some but not all phonological inventories. Their complexity, or difficulty, is due to independent properties of the sensorimotor system that are reflected in the grammar through these constraints. Marking statements may be active or deactivated. If a marking statement is active in a language, the marked configuration is not accepted in this language—it is illicit, and segments containing this configuration are absent from the language. It is obvious that if a feature combination is absent in a language, speakers of this language are never exposed to it. Under these conditions, speakers are unable to learn to coordinate the articulatory actions needed to implement that clustering of phonological properties. Therefore, when a marking statement is active, the targeted phonological configurations cannot be transformed into articulatory commands. It is assumed that structures that are not allowed are 'fixed up', or repaired by speakers. The repairs that occur in this case then indicate the featural/configurational manipulations that adjust the representations and make their conversion into articulatory commands possible (see Calabrese, 2005 for more discussion of this model).

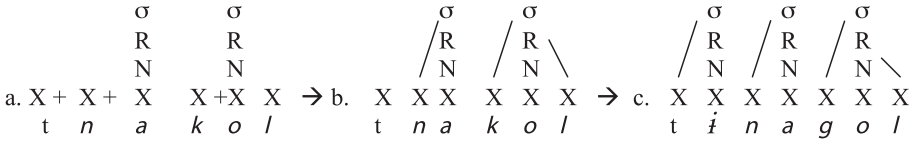
Romani and Calabrese (1998) shows how the phonological errors of an Italian aphasic patient, DB, can be accounted for by using the same markedness constraints and repairs that are independently needed in the phonological component of synchronic grammars. The pattern of errors in this patient involved syllabic configurations. Romani and Calabrese hypothesized that the level of syllabic complexity allowed by a speaker depends not only on the language he is speaking, but also on whether or not he has suffered brain damage. Brain damage can impair the ability to realize certain sequences of articulatory gestures and, therefore, can reset the degree of complexity allowed for syllabic configurations. Marking statements that are normally deactivated can be momentarily or permanently activated in aphasic patients, resulting in an increase of active negative constraints in aphasic patients in comparison with normal subjects. Patient DB had problems with hiatus configurations, i.e., sequences of vowels, and eliminated them in various ways: i) by deleting one of the vowel in hiatus (*empireo* → *empir_o)*, ii) by inserting a consonant between them *empireo* → *empirelo*, or iii) by resyllabifying the first vowel into a glide (*empireo* → *empiryo).* Now, although hiatus configurations are regularly allowed in Italian, they are disliked by many other languages. They can therefore be considered as being marked structures, and thus governed by a dedicated marking statement (*VV). This marking statement is deactivated in Italian. Romani and Calabrese (1998) argued that it had became active in DB's grammar and showed that his errors involving hiatuses could be accounted as instances of the repairs usually adopted across languages to eliminate these configurations: (i) deletion, (ii) consonant insertion and (iii) resyllabification.

A typical repair to fix illicit syllabic configurations is epenthesis, which involves the insertion of a syllabic nucleus to syllabify consonants that cannot be otherwise syllabified according to the constraints

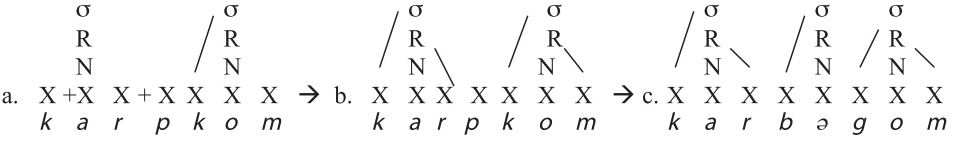
governing the syllable structure of a language. When morpheme concatenation creates a sequence of consonants that cannot be syllabified in onsets or codas, epenthesis inserts nuclei so that the sequence can be syllabified. For example, consider Lenakel (Blevins, 1995) where certain morphemes are represented by single consonants. These consonants are syllabified as onsets, with the exception of the word-final ones which become codas. In this language, no complex onsets and complex codas are allowed, i.e., the constraint *CCV, and *VCC are active. When syllabification would create violations of these constraints, epenthesis intervenes: /*t-n-ak-ol*/ → *tinágɔl* "you (sg.) will do it"; /*t-r-ep-ol*/ → *tirébɔl* "he will then do it"; /*kam- n- mãn-n*/ → *kàmnimánin* "for her brother"; /*k-ar-pkom*/ → *karbágɔm* "they are heavy".

Thus, given the morphological inputs in (4a), prevocalic consonants are syllabified as simple onsets and postvocalic ones in word-final position as codas. Otherwise, epenthesis occurs, as shown below (in (4/5a) we have the inputs; in (4/5b) the incorporation of simple onset and codas; in (4/5c) epenthesis):

(4)



(5)



Epenthesis as a strategy to repair complex syllabic configurations is also reported in the speech of English aphasic patients ((Buchwald, 2005): *bleed* → [bəlid], *glue* → [gəlu], *prune* → [pərun], *crab* → [kər b]). These examples can be analyzed by hypothesizing that the impairment in the patient has led to the activation of the constraint *CCV which blocks complex onsets (i.e., *blV, prV*). Syllabic configurations violating it are repaired by epenthesis as discussed above.

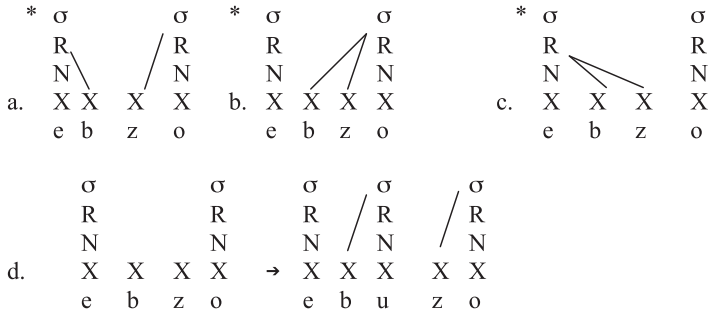
Let us turn to perception. When faced with a new word or morpheme, parity conditions must be met. In this case, the identification of the phonological structure of the exponent is the main goal of the perceptual process so that we can properly memorize it. A mental representation of the phonological structure of the exponents in terms of features must be constructed through analysis-by-synthesis. The hypothetical feature array extracted from the auditory representation generated from the acoustic signal is computed by applying to it the relevant phonological processes. The representation that is constructed through this derivation is the mental representation of the input word. The representation that is so derived is then converted into an auditory representation by computing the auditory correlates of acoustic effects of its articulatory implementation. If this auditory representation matches the auditory representation generated from the acoustic signal, the mental representation is accepted as the correct percept.

In the model outlined in this section, the difficulty that listeners have with 'foreign, unfamiliar' segments is accounted for in terms of an active marking statement: the foreign segments are blocked by the active marking statements of the listeners' language. Therefore, when faced with a new word containing an unfamiliar linguistic configuration, a foreign sound, listeners have an obvious problem because the input contains a configuration that is grammatically illicit—a configuration that is blocked by an active marking statement. In this case, the grammar prescribes the application of a repair to

adjust the configuration. The grammatical computation constructing the mental representation of a word with a foreign configuration must therefore include this repair. Hence when the hypothetical feature array extracted from the auditory representation generated from the acoustic signal of this word is computed, the illicit configuration contained there must be repaired. Application of the repair will generate an output representation that is grammatically correct, but mistaken.[8]

Japanese vowel epenthesis may therefore be accounted for in the following way. Japanese disallows complex onsets and complex codas. Simple codas are restricted to the first member of a geminate, or to a nasal glide. When faced with a word with such illicit clusters, a Japanese listener generates a preliminary hypothetical representation that contains a configuration that is illicit because of the active constraints characterizing his grammar. The listener resorts to epenthesis to fix this problem. Take the input [ebzo] for a Japanese listener. The consonantal sequence extracted from the signal cannot be organized into a syllable structure that is possible according to the grammatical constraints active in Japanese, as shown in (6a, b, and c). Hence it is repaired by epenthesis (6d).

(6)



The adjustments we observe in perception indicate that active constraints and repairs that are active in grammar as discussed above play a role in perception too.[9] So unfamiliar sound configurations that are disallowed by active constraints need to be repaired in perceptual representations. This repair results in a perceptual adaptation of the unfamiliar sounds. Parity cannot be reached in these cases insofar as the input acoustic signal contains an illicit configuration in L1. Construction of a grammatically licit representation is preferred to faithfulness to the input signal. A phonological illusion is therefore generated.

In this model, the difficulty that listeners have with 'foreign' sounds, in addition to foreign syllabic configurations, is also accounted for in terms of an active marking statement. It is assumed that marking statements govern the structure of phonemic systems. The absence of the configuration [−back, +round] in a language, i.e., of front rounded vowels [ü, ö], is formalized in terms of the marking statement *[−back, +round], which states that this combination is illicit in the language. Languages vary in what features combinations are allowed in their inventory. Calabrese (1988, 2005) argues that the following set of marking statements: (a) *[−low, −high], (b) *[−high, +ATR], (c) *[+low, −back], (d) *[−back, +round], (e) *[+high, −ATR], (f) *[+back, −round]/[__, −low], (g) *[+low, +round], (h) *

[8] This occurs if the marking statement remains active. However, through auditory exposure and motor training, the learner can also learn to produce the sound configuration, i.e. deactivate the marking statement. In this case the configuration will be perceived as normal.

[9] Observe that saying that perceptual adaptations involve a grammatical computation does not imply that the adaptations simply follow from the grammar of L1. There is evidence that the repairs found in perceptual adaptation may be different than those found in L1 grammar (Boersma & Hamann, 2009; Calabrese, 2009; Kabak & Idsardi, 2007). The input to the analysis and interpretation in linguistic perception is the word in its surface representation. Only the minimum that is necessary to fix the input is changed, the rest must be preserved. The preservation of the licit aspects of the input may explain why the treatment of L2 words in L1 often involve processes that are not part of L1 phonology (see Calabrese, 2009 for more discussion).

[+low, +ATR] accounts for the varying structure of vowel systems across languages. A language in which no marking statement is deactivated will have the vowel system /i, u, a/. Arabic is a language of this type. If a language deactivates the marking statement in (c), it will have the vowel system /i, u, æ, a/ . Latvian is a language of this type. If instead of (c), a language deactivates the marking statement in (a), it will have the vowel system /i, u, ɛ, ɔ, a/ which is found in Modern Greek, Spanish, Hawaiian, and many other languages. If, in addition to the marking statement in (a), a language also deactivates the marking statement in (b), it will have the vowel system /i, u, e, ɛ, o, ɔ, a/, which is found in standard Italian. If, instead, it deactivates the marking statements in (c) and (d), it will have the vowel system /i, y, u, ɛ, œ, ɔ, æ, a/, which is found in Finnish. The structure of other vowel systems can be accounted for in similar ways by marking-statement deactivation.

For a listener of a language with the vowel system /i, u, ɛ, ɔ, a/, the vowel [æ] is "foreign" because it is excluded by the marking statement *[+low, −back] which is active in his language. The same holds for the vowels [ü, ö] which are disallowed by the marking statement *[−back, +round] which is also active in his language.

As discussed above, if a marking statement is active in a language, the configuration marked by this statement is illicit in this language. Illicit configurations are fixed up by phonological repairs. A common repair adjusting illicit featural configurations involves deleting one of the illicit feature specifications and replacing it with the opposite specification. This type of repair can occur in production but crucially also in perception. For example, consider Italian listeners. The Italian vowel system does not have the [+low, −back] vowel /æ/ of the English word /kæt/ and Italians interpret this vowel as either [ɛ] or [a]. This can be explained as follows. The illicit configuration [+low, −back] of vowel [æ] may be fixed up by replacing [+low] with [−low]: [+low, −back] → [−low, −back], or by replacing [−back] with [+back]: [+low, −back] → [−low, +back]. In the first repair, the illicit vowel [æ] is replaced by vowel [ɛ], in the second one, the illicit vowel [æ] is replaced by vowel [a].

The other cases discussed in Section 2 can be analyzed in the same way. For example, the fact that Italian does not have the retroflex stop ʈ, ɖ found in Hindi means that the marking statement * [+distributed, −anterior] is active in Italian. This active marking statement triggers repairs in perception in Italian listeners so that the illicit configuration [+distributed, −anterior] of these consonants may be fixed up by replacing [−anterior] with [+anterior]: [ʈ, ɖ] → [t, d], or by replacing [−distributed] with [+distributed] (+subsequent affrication (see Calabrese, 2005: chap. 4): [tʃ, dʒ]).

7. Echoic memory

Given what we have discussed above, access to grammar, access to lexical information, and the ability to reconstruct, or even construct, phrases and sentences in the percept, in addition to grammatically induced phonological "illusions", indicate that perception clearly involves top-down processes. In this article, as already mentioned, I propose that a top-down system plays a fundamental role in speech perception. This system implements an analysis-by-synthesis of linguistic inputs.

The presence of these illusions indicates that our perception of linguistic reality, in particular of sounds and sound structure, is indirect, mediated by the L1 grammar. But then this reality in itself should not be perceivable. The input sound could never be apprehended without being interpreted in perception. This means that L2 sounds could not be learned. Also, given that the same grammar restrictions are characteristic of children, the consequence would be that they would never be able to acquire the language of the adults.

If what is perceived is always mediated through top-down knowledge, awareness of the acoustic shape of the given sound by a listener when his own pronunciation of this sound, i.e., his grammar, is different, should be impossible.

But this is contrary to our common experience of language. L2 sounds can be learned, albeit with difficulty, and children can learn their ambient language. Therefore there must be a component of the perceptual system that preserves acoustically faithful representations of the incoming signal so that novel representations can be constructed and learned and a new phonological system constructed.

In fact, it appears that more information is immediately available in auditory memory that can be reported. Neisser (1967) has called this auditory memory the echoic memory. It is clear that we need such a memory to process many aspects of speech information. Neisser (1967: 201) gives the example

of a foreigner who is told, "No, not zeal, seal." Foreigners would not be able to benefit from this information if they could not retain the 'z' long enough to compare it with the 's'.

Without the acoustic patterns stored in echoic memory the objective reality of the sounds to be learned could not be approximated. This is especially true for foreign sounds in second language acquisition, which can be distorted by phonological illusion, as we have seen in the preceding section. Only by assuming that their acoustic reality is stored in echoic memory can we account for their acquisition.[10] One can hypothesize that echoic memory preserves the high fidelity representation of the signal generated by the spectro-temporal receptive fields in core auditory cortex (Mesgarani et al., 2008). These are the auditory representations that are further analyzed in the successive, more abstract phases of speech perception.

Echoic memory as a part of the bottom-up component[11] of perception is necessary to account for language learning. According to Doupe and Kuhl (1999), for the child to learn to articulate the speech sounds in his or her linguistic environment, there must be the following mechanisms: (i) a mechanism by which sensory representations of speech uttered by others can be stored (echoic memory); (ii) a mechanism by which the sensory input is analyzed and converted into an articulatory representation (the analysis-by-synthesis component); and (iii) a mechanism by which the child's articulatory attempts can be compared against these stored representations, and by which the degree of mismatch revealed by this comparison can be used to shape future articulatory attempts (the 'comparator' of the analysis-by-synthesis component). Given that adults can learn foreign sounds, one must assume that such a network continues to operate throughout life (Houde & Jordan, 1998; Waldstein, 1989; Yates, 1963).

At this point it is important to deal with a common aspect of the speaker/listener's experience of foreign sounds that is aptly described in the following anecdote (from Jacobs & Gussenhoven, 2000: 203): "Valdman (1973) reports the reaction by a Haitian-Creole-speaking maid who attended evening literacy classes to her teacher's pronunciation of oeuf 'egg' as [ze]: she decided to leave the class. Although she herself pronounced it that way, she was aware that her bilingual employers realized it as [zø]". She did not know how to pronounce it, but she did know how it sounded. A listener can be aware of the acoustic shape of a given sound, although it is not present in his grammar. The same occurs in children. When my daughter was about two years old, she used to pronounce [ʃ] as [s]; therefore, she said [sIp] instead of [ʃIp] 'ship'. So once I tested her and, while pointing to the picture of a ship, I asked: "Is this a [sIp]?" Her answer: "No! it is a [sIp]!" Therefore, there must be a long-term memory system where auditory information on unfamiliar sounds must be preserved. It is in this long-term memory system of the Haitian maid that the auditory information about the standard sound [ø] for [zø] 'egg' was stored. Thus although this sound was not part of her grammatical competence, she could use that echoic memory representation for comparisons. The same for my two-year old daughter. She must have had long-term memory auditory information telling her that the initial sound of the English word for 'ship' was indeed [ʃ].

Now, echoic memory, like other sensory memories, is conceived as having a short temporal span lasting only for a brief period of time between 3 and 20 s according to Cowan (1984), Remez (2003). However, here we seem to require long-term memory of sensory auditory representations. Indeed, long-term traces of auditory representations are assumed in the literature on auditory representations (Näätänen, Paavilainen, Rinne, & Alho, 2007). Where is this auditory information about words of a language stored? As hypothesized by Chomsky (1986), there may be two different systems of linguistic knowledge, and therefore of long-term memory (see also Poeppel et al., 2008). One contains

[10] Observe that from the point of view of production, the acoustic patterns stored in echoic memory acquire the function of targets for speech motor control. We can then say that the motor implementation of feature representations, which characterize speech production, is constrained and under the control of the relevant acoustic targets. In this sense, the model outlined here is perhaps not too far away from the DIVA model of speech production of Guenther (1995, 2006), Perkell, Guenther, et al. (2004) and Perkell, Matthies, et al. (2004)—in which the auditory and acoustic consequences of the articulator movements play a fundamental role in production, although in a strictly computational and phonology-based framework.

[11] Following a model such as Trace (McClelland & Elman, 1986) (see also Klatt, 1980), but in a strictly computational formulation, perception must then contain both a bottom-up and a top-down component that run in parallel and interact with each other.

abstract, categorial, symbolic representations. This is the generative core of language that allows the production and perception of an infinite number of utterances. The other is token-based, containing representations closer to the signal, in which the experiential knowledge about language is encoded. The first system contains formal rules and constraints and the vocabulary that contains the lists of the exponents of the morphemes of the language; the other contains all types of linguistic information including token-based representations of signal (cf. exemplar models of memory (Goldinger, 1998, 2000; Goldstone & Kersten, 2003; Hintzman, 1986)). We can call this system the encyclopedia (see also Eco, 1997 about the necessity of such two systems of knowledge). A system storing knowledge about language is fundamental to account for the social dimension of the language. Listeners must be able to detect, know, and remember all types of phonetic detail about the speech of other members of the comunity in order to have adequate social interactions. So they must have memorized information about different types of speech registers, accent, dialects, and so on. One can hypothesize that auditory information about foreign sounds is stored in long-term memory echoic representations in the encyclopedia (see Section 9 for more discussion).[12]

8. Phonological deafening and perception

It is often stated that children become 'deaf' to foreign phonological contrasts in the process of language acquisition. Before six to nine months of age, infants apparently can discriminate any sounds that contrast in any language. By the end of the first year, however, it is stated that they apparently can no longer discriminate most sounds that do not contrast in the ambient language (Aslin, Jusczyk, & Pisoni, 1998; Best et al., 1995; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994; Werker & Lalonde, 1989; Werker & Tees,, 1984).

This developmental change is accounted for in language-learning models (cf. Best's Perception Assimilation model (Best, 1994, 1995), Flege's speech learning model (Flege, 1991), Kuhl's Native Language Magnet (Grieser & Kuhl, 1989; Iverson & Kuhl, 1995; Kuhl, 1991, 1993; Kuhl et al., 1992)) as a side effect of the infant having learned the phonological categories of the ambient language during this six to nine month period. It is proposed that by attracting both the ambient and foreign sounds the infant hears, these categories learned by 12 months deafen the child to differences it could detect six months earlier.[13] For example, adult speakers of languages with fixed stress (French, Finnish, Hungarian) are apparently less able to detect a shift in stress position in a word than a change in one of its segments (Dupoux et al., 1997; Dupoux, Peperkamp et al., 2001; Peperkamp & Dupoux,, 2002). Even highly fluent bilinguals appear to be deafened, too, if their exposure to the second language is too late. For example, some Spanish dominant Spanish-Catalan bilinguals who did not learn Catalan before 5–6 years of age apparently cannot discriminate Catalan contrasts not shared with Spanish such as high mid versus low mid vowels, / e, o / versus / ɛ, ɔ/, or voiced versus voiceless fricatives, / z / versus / s / (Pallier, Bosch, & Sebastian Gallés, 1997).

However, the notion of "phonological deafening" is quite problematic. Adult human beings can indeed hear foreign sounds, so for example Arabic pharyngeals can be easily detected by speakers of languages without pharyngeals like Italian or English. Furthermore, and more importantly, foreign sounds can be learned with appropriate training. If adults were really deafened to foreign sound categories, they would never be able to learn the sounds of a second language insofar as they would just be insensitive to them. The hypothesis of phonological deafness to non-native sounds cannot account for these facts. To solve this problem, one must assume that learners can indeed adequately detect the acoustic details––in term of cues and other acoustic patterns––of the non-native sounds, and preserve their 'aural' representation in an echoic memory before they are converted and interpreted in terms of articulatory representations. This auditory representation is the learning target that

[12] Anyone who has ever played with a short-wave radio knows that it is possible to understand whether the language of the tuned station is Chinese, Russian, Arabic, etc. without knowing these languages. In this case I assume that we are accessing an aural memory of how they are spoken contained in the Encyclopedia.

[13] This deafening is apparently temporary until apparently 5–6 years of age, when if the child does not get sustained exposure to the foreign language, it is assumed to become permanent.

must be analyzed and properly converted into a correct featural representation through cognitive adjustments.

Therefore, as Kingston (2003) observes, the deafening we observe in children cannot mean an inability to hear differences between acoustic stimuli but must rather refer to the weakening in the behavioral response to these differences. An infant is behaviorally deafened in this sense to foreign contrasts because he no longer responds differently when the stimulus changes from one member of a foreign contrast to the other. The infant can still hear the different acoustic stimuli, just does not respond to them as he would to native phonological contrast. They are linguistically not important for him.

Therefore as Werker has argued, 'developmental change does not involve loss' (Phillips, 2001; Werker, 1994; Werker & Logan, 1985). Earlier in Section 4, following Werker and Logan (1985), I introduced the distinction between phonemic and phonetic perception. When subjects perceive stimuli according to native-language phonological categories, they are demonstrating 'phonemic' perception. This is the normal mode of perception and this is the mode that causes 'deafening'. However, if necessary there is also a phonetic mode of perception. When subjects show a sensitivity to phonetic distinctions that are used in some other (not their native) languages, they are using phonetically relevant (or "phonetic") perception. It is the phonetic mode of perception that allows learning of non-native phonological categories.

The phonetic mode of perception does not immediately lead to learning the foreign segments. The phonetic mode—which is the mode needed to satisfy parity requirements—leads to analysis-by-synthesis and therefore to grammatical adjustments of foreign segments and configurations as discussed in Section 6. However, from this it does not follow that foreign sounds are impossible to detect acoustically. For adults, 'deafening' to non-native sounds does not imply inability to hear, to access the acoustic signal, but lack of ability to convert the landmark/cue pattern in the non-native sound into a licit phonological representation, and therefore to interpret them as instances of phonological contrasts. However, these acoustic patterns can be heard and can be actually preserved in echoic memory. With the appropriate articulatory training—namely, in terms of Calabrese (2005), if the learner deactivates the relevant grammatical constraints—the acoustic patterns stored in echoic memory can be converted into licit phonological representations, and the learner thereby acquires the non-native sound. Perception and production closely interact in the acquisition of the second language phonology as well as—obviously—in the acquisition of first-language phonology.

9. Conclusions

In the preceding sections I presented an assessment of some ideas, findings, and theories concerning speech perception organized in a hopefully scientifically coherent narrative. Thinking about these issues led me to some empirical and theoretical questions involving the relation between speech perception and the brain. Here I will consider some of the issues related to analysis-by-synthesis and briefly turn to the role of features in mental representations.

I begin by pointing out that in my view analysis-by-synthesis occurs only in perceptual tasks requiring parity. Therefore, it is not always needed in speech perception; as I discussed in Section 3, I hypothesize that exponents can be directly identified from auditory representations without recourse to analysis-by-synthesis.

Now, if we assume that the function of analysis-by-synthesis is achieving parity, it follows that the representations and computations found in this perceptual task must be the same as those found in production; otherwise parity could never be achieved. We then have two sets of representations and computations that are the same but belong to two different components, one to production and the other to analysis-by-synthesis. From the point of view of theoretical parsimony, it would be better to unify the two sets and say that (i) there is a single component that implements phonological computations, that (ii) this is located in the production components, and that (iii) analysis-by-synthesis accesses the production component. If there is an involvement of the production component, we expect activation of the neural area dedicated to production during analysis-by-synthesis. Thus, we expect the motor system to play a role when the parity requirements must be met. In Calabrese (2009), for example, I speculated that the phonological computation charactering analysis-by-synthesis could be implemented in the verbal working memory, the so-called phonological loop (Aboitiz & Garcia,

1997; Baddeley, 1992; Hickok & Poeppel, 2004), where phonological representations are kept active in the articulatory rehearsal component which involves left frontal cortices, portions of Broca's area and more dorsal pre-motor regions (Awh et al., 1996)). Evidence for access to production when parity conditions are required, specifically in the perception of foreign sounds, is given in Callan, Jones, Callan, and Akahane-Yamada (2004) who investigated with fMRI the neural processes underlying the perception of L2 phonemes by L2 learners. The same phonemes (i.e., /r/, /l/, or a vowel at the beginning of English syllables) were used for native English speakers and English-L2 speakers (i.e., low proficient Japanese-English bilinguals). Greater activity for second language over native-language speakers during perceptual identification of /r/ and /l/ relative to vowels was found in Broca's area, anterior insula, Wernicke's area, and parietal areas (including the supramarginal gyrus) while more extended activity for native-language speakers was found in the anterior STS.

The issue of mirror neurons should be addressed at this point. Recent neurological studies have argued that perceiving speech involves neural activity of the mirror neurons and the motor system. The mirror neurons are a particular class of neurons that exhibit excitations not only when an individual executes a particular action but also when the same individual observes the action being executed by another individual (see Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Rizzolatti & Craighero, 2004; Rizzolatti, Fogassi, & Gallese, 2001). Fadiga, Craighero, Buccino, and Rizzolatti (2002) argued that the same motor centers in the brain are active both in the production of speech and in speech perception, where the perceiver engages in no overt motor activity (see also Pulvermüller et al., 2006). Given what was proposed earlier, involvement of the production system and therefore of the motor system should occur only in the case of perceptual tasks involving analysis-by-synthesis. One could assume that these are the tasks used in the experimental conditions revealing the mirror neuron system. Since I hypothesized that analysis-by-synthesis occurs only when parity requirements must be met, in particular during sublexical tasks where the phonological structure of the incoming signal must be identified, mirror neuron activity should be restricted to these tasks.[14]

Involvement of the motor area, and of the production system, are therefore not necessary for speech perception in general, but only required when parity conditions must be met. In fact, the intact ability of Broca's aphasia patients for word recognition and comprehension despite widespread damage to the motor area of language, and therefore to the production system, shows that this is correct. As observed by Hickok, Buchsbaum, Humphries, and Muftuler (2003), Hickok and Poeppel (2007), the findings from brain injuries argue against the hypothesis that motor areas in the frontal lobe are always required for perception. One can hypothesize that under conditions that do not require parity, i.e., when only the meaning needs to be extracted from the input signal, exponents in the acoustic signal could be identified directly from the phonological sketches extracted from auditory representations of this signal without going through the analysis-by-synthesis component.[15]

On the other hand, Hickok and Poeppel (2007) argue for the existence of posterior cortical areas that are involved in the programming of production, in particular for a cortical field at the interface of the temporal and frontal lobes that, according to them, provides the critical substrate for mapping from input representations to output representations, and possibly could be the locus for analysis-by-synthesis. If analysis-by-synthesis is implemented in posterior cortical areas, one should not expect involvement of the motor areas even in tasks requiring parity conditions. In this case, the regeneration of the derivation of the input could be thought of in terms of an 'abstract' featural code, without activations in the motor areas. Future research must therefore decide whether the synthesis part of the

[14] Heim and Friederici (2003) have shown that although the production and perception of phonemes share the same neural network, there is a difference in the temporal dynamics of the components of this network: Wernicke's area shows a temporal dynamics primacy over Broca's area during perception while the reversed pattern occurs for producing sounds. Most importantly, Broca's area activates in perception only if a sublexical task like phonological segmentation is required. Heim and Friederici (2003: 2033) state: "Wernicke's area is activated first, and only if phonological segmentation is required, Broca's area is recruited. In production it appears that Broca's area is activated first in order to retrieve phonological information from the sound form store of the mental lexicon located in the pSTG [Wernicke's area]."

[15] According to Hickok and Poeppel (2007), auditory analysis is performed by the superior temporal gyrus (STG), construction of the phonological sketches by the superior temporal cortex, posterior STG, and the Superior Temporal Sulcus and finally computation of lexical representations by the middle temporal gyrus.

analysis-by-synthesis process involves an abstract computation in posterior cortical areas or concrete activations in frontal motor areas.

Notice now that although learning words, and any sublexical process requiring the analysis of phonological exponents, needs the mapping of acoustic invariant properties of the signal including these words into articulatory feature representations and subsequent analysis-by-synthesis, once a word is learned, used and heard many times, and therefore totally familiar, it is obviously uneconomical to go through analysis-by-synthesis every time that the same word is heard. One could assume, instead, that recognition of familiar and commonly used words and constructions may directly activate the exponents of the dictionary through the simple extraction of the phonological sketches from auditory representations generated by the input signal. One could therefore wonder if an analysis-by-synthesis system is relevant only for acquisition and for situations requiring parity, and after that everything is recomputed into overlearned templates.

Those I just discussed are not the only issues concerning analysis-by-synthesis that are worth considering. There are many others. Here I can consider only a few more.

In this article I focussed only on the analysis-by-synthesis procedures required for exponent identification, i.e., for the cases where phonological representations and computations are involved. However, similar analysis-by-synthesis procedures are needed for reconstructing the morpho-syntactic structure of the utterance and for the identification of its semantic and pragmatic content, i.e., for all of the tasks involved in language comprehension. Can we assume that multiple analysis-by-synthesis loops exist, mediating between different processing stages and levels of representation? Furthermore, if each 'level' of representation has its own analysis-by-synthesis process, how do they operate? Serially? Most probably they flow in parallel. In this case, how does the emerging output of each one affect the processing of the other levels?

I have proposed that analysis-by-synthesis can account for the generation of illusions in perception. In these cases, there is obviously a failure to achieve parity. Is there any general way to account for when such failures occur? Related to this is the more general issue: how is the match between the stored initial auditory representation and the output of the synthetic component implemented? What algorithm is used to achieve it? On what kind of similarity is it implemented? Is it a competitive process?

Analysis-by-synthesis can be divided into three subroutines: (i) the initial extraction of an hypothetical feature representation from initial auditory representations, (ii) the phonological computations/derivations from this abstract representation, and (iii) the comparator stages where the aforementioned match occurs. Can these subroutines be isolated using the tools of cognitive neuroscience? Furthermore, can they be manipulated so that we may understand their internal architectures?

We can now turn to some representational issues. I have argued that the phonological representations of exponents are constructed in terms of distinctive features. Features are most adequately conceived as abstract correlations between auditory patterns and articulatory commands. In this sense, they are abstract units of phonological computations and may remain abstract except during the last stages of articulatory implementation and during the initial stages of perception where they would involve concrete neural activations of the motor system and of the auditory system, respectively. With the exception of those peripheral stages, they would simply be computing units of abstract internal programs. Is this correct? Can it be shown? Are features abstract coding elements or do they always involve concrete neural activations in the motor and auditory systems? In this regard it is important to consider the issue of mirror neurons in perception. Earlier I proposed that activations in the motor area must be restricted to perceptual stages involving analysis-by-synthesis. One could then hypothesize that the motor activation associated with features occurs at the latest stages of the phonological computation in analysis-by-synthesis when after constructing the surface representations, the acoustic effects of their articulatory implementations are calculated. Is this correct?

In an approach that assumes features, sounds as unitary elements do not exist; sounds are nothing else than bundles of features. Perception of sounds is essentially perception of feature complexes. Therefore, sounds such as /e/, /a/, /ü/, /p/, /ʕ/ /ɣ/ etc. are nothing else than the relevant complexes of features, and their perception as unitary objects would be simply an illusion. One can then question the status of sounds in auditory representations, and in particular in the long-term echoic memory representations which I proposed in Section 7. Notice that these auditory representations must be quite abstract. Already an utterance is an abstract auditory object since the utterance must be extracted from

the ecologically chaotic acoustic input that arrives in our ears (Remez, 2003). Furthermore, although some behaviorally salient, important utterance may indeed be memorized, storage of the auditory representations of all utterances ever heard seems wasteful and psychologically dubious. Therefore the sound components of the utterances must be extracted to be memorized. This extraction obviously requires a major analytical process of abstraction and segmentation. First of all, all linguistically irrelevant properties, such as the voice characteristics of the speaker who uttered the word, its rate of speech, distortions caused by a cold or sore throat, and so on must be removed from the input signal. Secondly, and most importantly, morphological and phonological segmentation must be implemented in the formation of the memorized linguistic representation of the sound. The resulting representation is necessarily abstract.

Since these auditory representations are in any case abstract, it is plausible to assume that they may have a format resembling those of higher cortical areas where distinctive features are used. It is in fact widely accepted that the primary auditory cortex responds distinctively to stimuli that are linguistically meaningful, in particular to acoustic distinctions that are correlated with distinctive features (e.g. Griffiths & Warren, 2002, 2004; Jacquemot, Pallier, LeBihan, Dehaene, & Dupoux, 2003; Näätänen, 2001; Näätänen, Tervaniemi, Sussman, Paavilainen, & Winkler, 2001; Nelken, Fishbach, Las, Ulanovsky, & Farkas, 2003; Obleser, Scott, & Eulitz, 2006; Scott, 2005; Shamma, 2008; Wang, Lu, Snider, & Liang, 2005) and that there are subpopulations of neurons that respond well to the distinctive acoustic features of a given phoneme and hence can fully analyze that phoneme in its featural complexity (Mesgarani et al., 2008).[16] Therefore the long-term echoic memory representations of sounds may already be analyzed in terms of features. They would be phonetically faithful to the input, although they may also be phonologically illicit in the case of foreign sounds that would contain featural configurations not accepted in the grammar of L1. An issue is the preservation of linguistically irrelevant details of speech like auditory information about the finer positions of the vocal tract during sound production in a given language. Are these details stored in long-term echoic memory representations? And how do speakers learn about them? Perhaps they are accessed during short-term echoic memory where through analysis-by-synthesis, they can lead to refinements of the processes of articulatory implementations. Can this be shown?

I believe that these are some of the questions that the joint interdisciplinary efforts of the theoretical modeling of cognitive capacities, the study of pathology, and brain imaging techniques will have to answer in the future. Many more questions can be asked, but I will stop here for the moment.

References

Aboitiz, F., & Garcia, V. R. (1997). The evolutionary origin of language areas in the human brain. A neuroanatomical perspective. *Brain Research Reviews, 25*, 381–396.

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology: Cognition, perception, and language* (pp. 147–254). New York: Wiley.

Awh, E., Jonides, J., Smith, E. E., Schumacher, E. H., Koeppe, R. A., & Katz, S. (1996). Dissociation of storage and rehearsal in working memory: PET evidence. *Psychological Science, 7*, 25–31.

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556–559.

Bach, E., & Harms, R. (1972). How do languages get crazy rules? In R. Sockwell, & R. Macaulay (Eds.), *Linguistic change and generative theory* (pp. 1–21) Bloomington: Indiana University Press.

Best, C. T. (1994). The emergence of nativelanguage phonological influences in infants: A perceptual assimilation hypothesis. In J. C. Goodman, & H. C. Nusbaum (Eds.), *The development of speech perception* (pp. 167–224). Cambridge, MA: MIT Press.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange, & J. J. Jenkins (Eds.), *Cross-language speech perception* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., Lafleur, R., & Silverisenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative speech contrasts. *Infant Behavior and Development, 18*, 339–350.

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

Blevins, J. (1995). The syllable in phonological theory. In J. Goldsmith (Ed.), *The Handbook of phonological theory* (pp. 206–244). Oxford: Blackwell.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience, 8*, 389–395.

[16] Sensitivity to sounds as unitary objects in this case would be created by groups of neurons with very high selectivity to specific acoustic features.

Boersma, P., & Hamann, S. (2009). Loanword adaptation as first-language phonological perception. In A. Calabrese, & L. Wetzels (Eds.).
Borsley, R. D. (1997). Mutation and case in Welsh. In E. Guilfoyle (Ed.), *Canadian Journal of Linguistics, 42(Special issue: Topics in Celtic Syntax)* (pp. 31–56).
Borsley, R. D. (1999). Mutation and constituent structure in Welsh. *Lingua, 109*, 267–300.
Bromberger, S., & Halle, M. (1992). The ontology of phonology. In S. Bromberger (Ed.), *What we know we don't know* (pp. 209–228). Chicago: Chicago University Press.
Bromberger, S., & Halle, M. (1997). The content of phonological signs: a comparison between their use in derivational theories and in optimality theories. In I. Roca (Ed.), *Derivations and constraints in phonology* (pp. 93–122). Oxford: Oxford University Press.
Bromberger, S., & Halle, M. (2000). The ontology of Phonology (revised). In N. Burton-Roberts, P. Carr, & G. Docherty (Eds.), *Phonological knowledge. Conceptual and empirical issues* (pp. 19–37). Oxford: Oxford University Press.
Brown, C. (1998). The role of the Ll grammar in the L2 acquisition of segmental structure. *Second Language Research, 14*, 136–193.
Brown, C. (2000). The interrelation between speech perception and phonological acquisition from infant to adult. In J. Archibald (Ed.), *Second language acquisition and linguistic theory* (pp. 4–63). Oxford: Blackwell.
Browman, C. (1980). Perceptual processing: evidence from slips of the ear. In V. Fromkin (Ed.), *Errors in linguistic Performance: Slips of the tongue, ear, pen, and hand*. New York: Academic Press.
Buchwald, A. (2005). Representing sound structure: evidence from aphasia. In J. Alderete, et al. (Eds.), *Proceedings of the 24th West Coast Conference on formal linguistics* (pp. 79–87). Somerville, MA: Cascadilla Proceedings Project.
Calabrese, A. (1988). *Towards a theory of phonological alphabets*. Ph. D. diss., Massachusetts Institute of Technology.
Calabrese, A. (1995). A constraint-based theory of phonological markedness and simplification procedures. *Linguistic Inquiry, 26*(2), 373–463.
Calabrese, A. (2005). *Markedness and economy in a derivational model of phonology*. Berlin: Mouton-De Gruyter.
Calabrese, A. (2009). Perception, production and acoustic inputs in loanword phonology. In A. Calabrese, & L. Wetzels (Eds.).
Calabrese, A. (2010). Metaphony in romance. To appear in the Blackwell companion in phonology. To appear. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell companion in phonology*. Oxford: Blackwell.
Calabrese, A., & Wetzels, L. (Eds.). (2009). *Studies in loan phonology*. Amsterdam: John Benjamins.
Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York, NY: Praeger.
Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.
Callan, D., Jones, J., Callan, A., & 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.
Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1*, 185–195.
Clements, G. N. (2009). The role of features in phonological inventories. In E. Raimy, & C. E. Cairns (Eds.), *Contemporary views on architecture and representations in phonology* (pp. 19–69). Cambridge: the MIT Press.
Clements, G. N., & Keyser, S. J. (1983). *CV phonology: A generative theory of the syllable*. Cambridge, MA: Massachusetts Institute of Technology Press.
Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin, 96-2*, 341–370.
Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport, 8*(4), 919–924, 3.
Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience, 12*, 635–647.
Delgutte, B., & Kiang, N. Y. S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *Journal of Acoustical Society of America, 75*, 897–907.
Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179.
Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research, 91*, 176–180.
Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience, 22*, 567–631.
Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J. (1997). A distressing "deafness" in French? *Journal of Memory and Language, 36*, 406–421.
Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1568–1578.
Dupoux, E., Pallier, C., Kakehi, K., & Mehler, J. (2001). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes, 16*, 491–505.
Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress 'deafness'. *Journal of the Acoustical Society of America, 110*, 1606–1618.
Eco, U. (1997). *Kant and the Platypus. Essays on language and cognition*. New York: Harcourt Brace.
Eulitz, C., & Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience, 16*, 577–583.
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.
Flege, J. E. (1991). Perception and production: The relevance of phonetic input to L2 phonological learning. In C. Ferguson, & T. Huebner (Eds.), *Crosscurrents in second language acquisition and linguistic theories*. Philadelphia, PA: John Benjamins.
Fowler, C. A. (1986). An event approach to the study of speech perception from a direct- realist perspective. *Journal of Phonetics, 14*, 3–28.
Garnes, S., & Bond, Z. (1980). A slip of the ear: A snip of the ear? A slip of the year? In I. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* New York: Academic Press.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. In A. Cutler, J. M. McQueen, & R. Zondervan (Eds.), *Proceedings of the workshop on Spoken Word Access Processes (SWAP)* (pp. 155–159). Nijmegan: Max Plack Institute for Psycholinguistics.
Goldstone, R. L., & Kersten, A. (2003). Concepts and categorization. In A. Healy, & R. Proctor (Eds.), *Comprehensive handbook of psychology*. New Jersey: Wiley.
Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences, 25*, 348–353.
Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*, 885–890.
Grieser, D., & Kuhl, P. (1989). Categorization of speech by infants: support for speech- sound prototypes. *Developmental Psychology, 25.4*, 577–588.
Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102*, 594–621.
Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders, 39*(2006), 350–365.
Halle, M. (1995). Feature geometry and feature spreading. *Linguistic Inquiry, 26*, 1–46.
Halle, M. (2002). *From memory to speech and back. Papers on phonetics and phonology 1954–2002*. Berlin: Mouton De Gruyter.
Halle, M., & Stevens, K. (1962). Speech recognition: a model and a program for research. *IEEE Transactions on Information Theory, 8*, 155–159, Now in Halle (2002).
Halle, M. & Stevens, K (1971). A note on laryngeal features. MIT Quarterly Progress Report 101, 198–212.
Halle, M., & Stevens, K. (1991). Knowledge of language and the sounds of speech. In J. Sundberg, et al. (Eds.), *Music. Language, speech and brain* (pp. 1–91). London: McMillan Press, Now in Halle (2002) 176–95.
Halle, M., Vaux, B., & Wolfe, A. (2000). On feature spreading and the representation of place of articulation. *Linguistic Inquiry, 31*, 387–444.
Hawkins, S. (2009). Phonological features, auditory objects and illusions. *Journal of Phonetics*, .
Heim, S., & Friederici, A. D. (2003). Phonological processing in language production: time course of brain activity. *Neuroreport, 14*, 2031–2033.
Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience, 15*, 673–682.
Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*, 67–99.
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech perception. *Nature Reviews Neuroscience, 8*, 393–402.
Hintzman, D. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Review, 93*, 411–428.
Houde, J., & Jordan, M. F. (1998). Sensorimotor adaptation in speech production. *Science, 20*(279), 1213–1216.
Hwang, S.-O., Monahan, P. J., & Idsardi, W. J. (2010). Underspecification and asymmetries in voicing perception. *Phonology, 27*, 205–224.
Idsardi, W. J. (2007). Some MEG correlates for distinctive features. *Proceedings of 16th International Congress of Phonetic Sciences*, .
Ingram, J., & See-Gyoon, P. (1998). Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America, 103*, 1161–1174.
Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of Acoustical Society of America, 97*, 553–562.
Jacobs, H., & Gussenhoven, C. (2000). Loan phonology: perception, salience, the lexicon and OT. In J. Dekkers, F. van der Leeuw, & J. van de Weijer (Eds.), *Optimality theory. Phonology, syntax, and acquisition* (pp. 193–210). Oxford: Oxford University Press.
Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. Acoustics Laboratory. MIT, Technical Report No. 13. (Republished 1967, 7th edn., Cambridge, MA: MIT Press).
Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *Journal of Neuroscience, 23*(29), 9541–9546.
Kabak, B., & Idsardi, W. J. (2007). Speech perception is not Isomorphic to phonology: the case of perceptual epenthesis. *Language and Speech, 50*, 23–52.
Kang, Y. (2010). Loanword phonology. To appear. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell companion in phonology*. Oxford: Blackwell.
Kazanina, N., Phillips, C., & Idsardi, W. J. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences, 103*, 11381–11386.
Kenstowicz, M. (1994). *Phonology in generative grammar*. Cambridge: Blackwell.
Kenstowicz, M., & Sohn, H.-S. (2001). Accentual adaptation in North Kyungsang Korean. In M. Kenstowicz (Ed.), *Ken Hale. A life in language* (pp. 239–270). Cambridge, MA: MIT Press.
Kim, H. (2009). Korean adaptation of English affricates and fricatives in a feature-driven model of loanword adaptation. In A. Calabrese, & L. Wetzels (Eds.).
Kingston, J. (2003). Learning foreign vowels language and speech. *Language and Speech, 46*(2–3), 295–349.
Klatt, D. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*, 93–107.
Kuhl, P. K. (1993). Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In B. de Boysson Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 259–274). Dordrecht, Netherlands: Kluwer Academic Publishers.
Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by six months of age. *Science, 255*, 606–608.
LaCharité, D., & Paradis, C. (2005). Category preservation and proximity vs. phonetic approximation in loanword adaptation. *Linguistic Inquiry, 36*, 223–258.
Ladefoged, P., DeClerk, J., Lindau, M., & Papcun, G. (1972). An auditory-motor theory of speech production. *UCLA Working Papers in Phonetics, 22*, 48–75.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. In C. Gussenhoven, & N. Warner (Eds.), *Lab phon 7* (pp. 637–676). Berlin: Mouton.
Lahiri, A., & Reetz, H. (2010). Distinctive features: phonological underspecification in representation and processing. *Journal of Phonetics, 38-1*, 44–59.
Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: Bradford Books.
Liberman, A. M., Cooper, E., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74.6*, 431–461.
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory or speech perception revised. *Cognition, 21*, 1–36.
Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243*, 489–494.
Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*, 187–196.
Malecot, A. (1960). Vowel nasality as a distinctive feature in American English. *Language, 36*, 222–229.
Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71–102.
Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.
Matthews, J., & Brown, C. (2004). When Intake exceeds input: language specific perceptual illusions induced by L1 prosodic constraints. *International Journal of Bilingualism, 8-1*, 5–27.
McCarthy, J. (1983). Consonantal morphology in the Chaha verb. In M. Barlow, D. Flickinger, & M. Wescoat (Eds.), *Proceedings of the West Coast Conference on formal linguistics 2*. Stanford, California: Stanford Linguistics Association.
McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.
Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, SA. (2008). Phoneme representation and classification in primary auditory cortex. *Journal of Acoustical Society of America, 123*(2), 809–909.
Mielke, J. (2008). *The Emergence of distinctive features*. Oxford: Oxford University Press.
Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition, 84*, 55–71.
Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent MMNm. *Psychophysiology, 38*, 1–21.
Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology, 118*, 2544–2590.
Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). 'Primitive intelligence' in the auditory cortex. *Trends in Neurosciences, 24*(5), 283–288.
Nearey, T. M. (1980). On the physical interpretation of vowel quality: cinefluorographic and acoustic evidence. *Journal of Phonetics, 8*, 213–241.
Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
Nelken, I. (2008). Processing of complex sounds in the auditory system. *Current Opinion in Neurobiology, 18*, 413–417.
Nelken, I., Fishbach, A., Las, L., Ulanovsky, N., & Farkas, D. (2003). Primary auditory cortex of cats: feature detection or something else? *Biological Cybernetics, 89*(5), 397–406.
Obleser, J., Scott, S. K., & Eulitz, C. (2006). Now you hear it, now you don't: Transient traces of consonants and their non speech analogues in the human brain. *Cerebral Cortex, 16*, 1069–1076.
Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of Acoustical Society of America., 99*, 1718–1725.
Paavilainen, P., Jaramillo, M., Näätänen, R., & Winkler, I. (1999). Neuronal populations in the human brain extracting invariant relationships from acoustic variance. *Neuroscience Letters, 265*, 179–182.
Paavilainen, P., Simola, J., Jaramillo, M., Näätänen, R., & Winkler, I. (2001). Preattentive extraction of abstract feature conjunctions from auditory stimulation as reflected by the mismatch negativity (MMN). *Psychophysiology, 38*, 359–365.
Pallier, C., Bosch, L., & Sebastian Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition, 64*(3), B9–B17.
Paradis, C. (1988). On constraints and repair strategies. *The Linguistic Review, 6*, 71–96.
Peperkamp, S., & Dupoux, E. (2002). A typological study of stress "deafness". In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology 7* (pp. 203–240). Berlin: Mouton de Gruyter.
Peperkamp, S., & Dupoux, E. (2003). Reinterpreting loanword adaptations: the role of perception. In *Proceedings of the I5th International Congress of phonetic sciences* (pp. 367–370).
Peperkamp, S., Vendelin, I., & Nakamura, K. (2008). On the perceptual origin of loanword adaptations: experimental evidence from Japanese. *Phonology, 25*, 129–164.
Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., & Tiede, M. (2004). Cross-subject correlations between measures of vowel production and perception. *Journal of the Acoustical Society of America, 116*(4 Pt. 1), 2338–2344.
Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., & Marrone, N. (2004). The distinctness of speakers'/s-sh/contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research, 47*, 1259–1269.
Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cognitive Science, 25*, 711–731.
Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience*, .
Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science, 21*(5), 679–693.
Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*, 245–255.
Poeppel, D., & Idsardi, W. J. (2010). *Recognizing words from speech: the perception-action-memory loop*. To appear in Festschrift for XXX.
Poeppel, D., Idsardi, W., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society London B, 363*, 1071–1086.
Poeppel, D., & Monahan, P. J. (2010). *Feedforward and feedback in speech perception: Revisiting analysis-by-synthesis*. Language and Cognitive Processes First published on: 01 July 2010 (iFirst).
Polivanov, E. D. (1931). *La perception des sons d'une langue »trangÀre. Travaux du Cercle Linguistique de Prague 4.79–96*. [*English translation, The subjective nature of the perceptions of language sounds, selected works: articles on general linguistics, pp. 223–237. The Hague: Mouton, 1974*].

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *J. Experimental Psychology, Human Perception, and Performance, 20*, 421–435.
Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Malden, MA, Oxford, UK: Blackwell.
Pulvermüller, F., Huss, M., Kheri, 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, 103*, 7865–7870.
Remez, R. E. (2003). Establishing and maintaining perceptual coherence: unimodal and multimodal evidence. *Journal of Phonetics, 31*.
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192.
Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*, 661–670.
Riordan, C. J. (1977). Control of vocal-tract lengthin speech. *Journal of Acoustical Society of America, 62*, 998–1002.
Romani, C., & Calabrese, A. (1998). Syllabic constraints in the phonological errors of an aphasic patient. *Brain and Language, 64*, 83–121.
Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2*, 437–442.
Samuel, A. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474–494.
Sapir, E. (1933). The psychological reality of phonemes. In Mandelbaum. (Ed.), *Selected Writings of Edward Sapir in language, culture and personality* (pp. 46–60). Berkeley: University of California Press, 1949.
Scott, S. K. (2005). Auditory processing–speech, space and auditory objects. *Current Opinion in Neurobiology, 15*, 197–201.
Shamma, S. (2008). On the emergence and awareness of auditory objects. *PLoS Biology, 6*(6), e155.
Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., et al. (2002). Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. *Neuroreport, 13*(14), 1813–1816.
Silverman, D. (1992). Multiple scansions in loanword phonology: evidence from Cantonese. *Phonology, 9*, 289–328.
Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In P. B. Denes, & E. E. David, Jr (Eds.), *Human communication: A unified view* (pp. 51–66). New York: McGraw-Hill.
Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3–46.
Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.
Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of Acoustical Society of America, 111*(4).
Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1999). Linear correlates in the speech signal: the orderly output constraint. *Behavioral and Brain Sciences, 21*, 287–299.
Sussman, E., Winkler, I., Huotilainen, M., Ritter, W., & Näätänen, R. (2002). Top-down effects can modify the initially stimulus-driven auditory information. *Cognitive Brain Research, 13*, 393–405.
Tervaniemi, M., Saarinen, J., Paavilainen, P., Danilova, N., & Näätänen, R. (1994). Temporal integration of auditory information in sensory memory as reflected by the mismatch negativity. *Biological Psychology, 1994*(38), 157–167.
Trubetzkoy, N. (1939). *Grundzuge der Phonologie*. English translation by C. A. M. Baltaxe. Berkeley: University of California Press, 1969. Gottingen: Vandenhoeck and Ruprecht.
Valdman, A. (1973). Some aspects of decreolization in Creole French. In T. Sebeok (Ed.), *Diachronic, areal, and typological linguistics. Current trends in linguistics, vol. II*. The Hague: Mouton.
Vaux, B. (2010). Feature analysis. In Patrick Hogan (Ed.), *Cambridge encyclopedia of the language sciences*. Cambridge: Cambridge University Press.
Vitevitch, M. S., & Luce, P. A. (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science, 9*, 325–329.
Waldstein, R. S. (1989). Effects of postlingual deafness on speech production: implications for the role of auditory feedback. *Journal of the Acoustical Society of America, 88*, 2099–2144.
Wang, X., Lu, T., Snider, R. K., & Liang, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature, 435*, 341–346.
Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392–393.
van Wassenhove, V., Grant, K., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *National Academy of Sciences: National Academy of Sciences USA, 102*, 1181–1186.
Werker, J. F. (1994). Cross-language speech perception: developmental change does not involve loss. In H. C. Nusbaum, & J. Goodman (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 93–120). Cambridge, MA: MIT Press.
Werker, J. F., & Lalonde, C. E. (1989). Cross-language speech perception: initial capabilities and developmental change. *Developmental Psychology, 24*, 672.
Werker, J. F., & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics., 37*, 35–44.
Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization. *Infant Behavior and Development, 7*, 49–63.
Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin, 60*, 213–251.