Chapter 1

The dynamical approach to speech perception: From fine phonetic detail to abstract phonological categories

Much research has been devoted to exploring the representations and processes employed by listeners in the perception of speech. There is in that domain a longstanding debate between two opposite approaches. Abstractionist models, on the one hand, are based on the assumption that an abstract and speaker-independent phonological representation is associated with each word in the listener's mental lexicon. In exemplar models of speech perception, on the other hand, words and frequently-used grammatical constructions are represented in memory as large sets of exemplars containing fine phonetic information. In the present paper, the opposition between abstractionist and exemplar models will be discussed in the light of recent experimental findings that call this opposition into question. In keeping with recent proposals made by other researchers, we will argue that both fine phonetic detail and abstract phonological categories are likely to play an important role in speech perception. A novel, hybrid approach, that aims to get beyond the abstraction vs. exemplar dichotomy and draws on the theory of nonlinear dynamical systems, as applied to the perception of speech by Tuller et al. (1994), will be outlined.

1. Two approaches to the perception of speech

In spite of the huge variability shown by the speech signal both within and across speakers, listeners are in most situations able to identify spoken words and get access to their meaning in an effortless manner. A major challenge in studies of speech perception is to understand how such variations in the pronunciation of words are dealt with by listeners in the sound-to-meaning mapping. There is still much disagreement about the nature of the processes and representations that this mapping may involve.

According to proponents of the highly influential abstractionist approach, the speech signal is converted by listeners into a set of context-independent

abstract phonological units, in a way that leads variations in how a given word is produced to be factored out at an early stage of processing (e.g., Fitzpatrick and Wheeldon, 2000; Lahiri and Reetz, 2002; Stevens, 2002). It is often hypothesized that inter-individual variations are removed prior to building up this abstract representation by means of a speaker normalization procedure (see Johnson, 2005b, for a recent review). A clear demarcation is established in that framework between the surface phonetic form of a word and the underlying phonological representation associated with that word. The abstractionist approach contends that the phonological representation for each word is both unique and permanently stored in memory. Readers are referred to Cutler et al. (2007), Eulitz and Lahiri (2004), Lahiri and Reetz (2002), Pallier et al. (2001), and Stevens (2002), for recent important papers in the abstractionist framework.

In the terminology employed by Pitt and Johnson (2003), abstractionist models such as Lahiri and Reetz's (2002) FUL model offer a representation-based solution to the speech variability problem. According to the FUL (Featurally Underspecified Lexicon) model, phonological representations in the lexicon are underspecified for certain features such as [coronal], and it is assumed that this makes listeners insensitive to surface variations shown by words for these features. For example, assimilation of word-final coronals to the place of articulation of the following consonant is considered to remain consistent with the underspecified phonological representation of the carrier word and, because of this so-called no-mismatch relationship between the word's surface and underlying forms, assimilation is expected to have no disruptive effect on the identification of the word. As opposed to representation-based models, processing-based models rely on the assumption that recognizing the assimilated variant of a word entails recovering the word's unassimilated shape through a phonological inference process.

In spite of the fact that they too are representation-oriented, exemplar models stand, in their most prototypical form at least, in sharp opposition to abstractionist models like FUL along many dimensions. A major difference with the abstractionist stance relates to the fact that each exemplar is viewed as corresponding to a language chunk that is stored in memory with all the details specific to the particular circumstances in which it has been produced or encountered. This includes sensory-motor, semantic and pragmatic characteristics, but also indexical information about the speaker's identity and the situation of occurrence, to mention but a few properties. Exemplars are therefore deeply anchored into their context of occurrence in the largest possible

sense, and this has drastic implications for how spoken language may be represented in the speaker's brain. In some exemplar-based theories, this is at variance with viewing linguistic utterances as being built from a pre-defined set of context-independent phonological primitives. In a radical departure from this widely accepted combinatorial view to language, Bybee and McClelland (2005) extend sensitivity to context and non-uniformity to all levels of linguistic analysis, and go as far as claiming that "there is no analysis into units at any level or set of levels that will ever successfully and completely capture the realities of synchronic structure or provide a framework in which to capture language change".

As pointed out by Johnson (2005a), the exemplar approach to sensory memory is well established in cognitive psychology for more than a century. Goldinger (1996, 1998) and Johnson (1997b) drew on this general theoretical framework to develop the first exemplar-based models of speech perception (see Pierrehumbert, 2006, for a historical overview), although these models have a number of prominent recent precursors: Klatt's (1979) Lexical Access from Spectra model, and Elman's (1990) Recurrent Neural Network, for example, were also based on the assumption that lexical representations are phonetically rich. Exemplar models today present a major challenge to the better-established abstractionist approach, with far-reaching implications not only for phonetics and phonology, but also and more generally for our understanding of language structure and language use (e.g., Barlow and Kemmer, 2000). In their current form, exemplar models of speech perception also raise a number of important theoretical and empirical questions, and it is on some of these questions that we focus in the following section.

2. Roles of fine phonetic detail and abstract phonological categories in speech perception

In recent years, an increasing body of research has been conducted on the listener's sensitivity to properties of the speech signal that are generically referred to as "fine phonetic detail" (FPD, hereafter). This research suggests that FPD has a significant impact on speech perception and understanding, in some circumstances at least. FPD includes allophonic variation, sometimes specific to certain words or classes of words (Pierrehumbert, 2002), as well as sociophonetic variation, broadly construed as being associated with the speaker's identity, her/his age, gender, and social category (Foulkes and Docherty, 2006). Fine phonetic detail is designated as such in the sense that

it is to be distinguished from the local and most perceptually prominent cues associated with phonemic contrasts in the speech signal (Hawkins, 2007). It may, however, encompass substantial acoustic variations, such as those that exist between male and female voices. FPD is therefore identified as "detail" with respect to a specific theoretical viewpoint, namely the traditional segmental approach to speech perception and production. A form of antiphrasis, it refers to phonetic properties that are judged non-essential in the identification of speech sounds in a theoretical framework whose limits the exemplar approach endeavors to demonstrate. The goal of current research on FPD is to show that FPD *is* important in speech perception, and, therefore, that a change of theoretical perspective is called for.

Recent studies have provided evidence that perceptually-relevant allophonic variation includes vowel-consonant acoustic transitions (e.g., Marslen-Wilson and Warren, 1994), within-category variations in voice onset time (Allen and Miller, 2004; Andruski et al., 1994; McMurray et al., 2002, 2003), long-domain resonance effects associated with liquids (West, 1999), and graded assimilation of place of articulation in word-final coronals (e.g., Gaskell, 2003; the studies cited here were conducted on either American or British English). To a certain extent, however, the listener's sensitivity to allophonic variation was established much earlier. For example, studies conducted in the 1970s and 1980s consistently showed that coarticulation between neighboring segments provides listeners with perceptually-relevant cues to segment identity. A well-known example is regressive vowel-to-vowel coarticulation in English, which allows the identity of the second vowel to be partly predictable from the acoustic cues associated with it in the first vowel (Martin and Bunnell, 1981, 1982). In addition to allophonic variation, it has been repeatedly demonstrated in recent years that listeners are sensitive to the individual characteristics of the speaker's voice in word recognition (e.g., Goldinger, 1996, 1998). Although allophonic and between-speaker variation are often put together under the FPD generic term, note that there is evidence suggesting that these two types of phonetic variation are dealt with in different ways by listeners (Luce and McLennan, 2005).

Further empirical evidence for the role of FPD in speech perception has recently arisen from a growing number of studies centered on the phonetics of conversational interaction (e.g., Couper-Kuhlen and Ford, 2004). A major issue of interest in these studies is the tendency shown by participants in a conversation to imitate each other. Imitation seems to occurs at every level of the conversational exchange, and that includes the phonetic level (Giles et al.,

1991). For example, Pardo (2006) had different talkers produce the same lexical items before, during and after a conversational interaction, and found that perceived similarity in pronunciation between talkers increased over the course of the interaction and persisted beyond its conclusion. Phonetic imitation, or phonetic convergence, is a mechanism that may be actively employed by talkers to facilitate conversational interaction by contributing to set a common ground between them. Phonetic imitation is also known to play a central role in speech development in infants (Goldstein, 2003; Meltzoff and Moore, 1997), and it has been assumed in recent work to be one of the key mechanisms that underlie the emergence and evolution of speech sound systems (e.g., de Boer, 2000). The behavioral tendency shown by humans to imitate others may be connected at the brain level with the presence of mirror neurons, whose role in in the production, perception and acquisition of speech now seems well established (Studdert-Kennedy, 2002; Vihman, 2002). Crucially for the present paper, phonetic convergence demonstrates that listeners are sensitive to speaker-dependent phonetic characteristics, which have an influence on both the dynamics of conversational interaction, and across a longer time range the representations associated with words in memory, when that interaction has ended. Such sensivity to context in listeners has led researchers like Tuller and her colleagues to contend that speech perception studies should focus on the listener's individual behavior in its situation of occurrence, as opposed to abstract linguistic entities ("the focus of the analysis must be the individual, not the language", Case, Tuller & Kelso, 2003, see also Tuller, 2004).

It is important to point out that by itself, the fact that listeners are sensitive to FPD is not inconsistent with abstractionist models of speech perception. For example, Stevens (2004) contends that in addition to what he refers to as the defining articulatory/acoustic attributes associated with distinctive features (e.g., for stop consonants, the spectrum of the release burst), the phonetic implementation of these features involves so-called language-specific enhancing gestures, which allow the features' perceptual saliency to be strenghtened (e.g., tongue-body positioning, for tongue-blade stop consonants, see Stevens, 2004). Although enhancing gestures can be regarded as fine phonetic detail, according to how this term is defined above, they are attributed an important role in Stevens' abstractionist model of lexical access (Stevens, 2002). Likewise, the TRACE model of speech perception (McClelland and Elman, 1986) relies on the assumption that fine-grained acoustic properties may have an impact on word recognition, although TRACE too

may be regarded as an abstractionist model (it contains an infralexical phonemic level of processing and, at the lexical level, each word is represented by a single processing unit). TRACE accounts for part of the listener's sensitivity to FPD by modelling the flow of acoustic information within the speech processing system by means of a set of continuous parameters. It is also designed to explain how fine-grained coarticulatory variation is taken into account in the on-line identification of phonemes (Elman and McClelland, 1988). Thus, the assumption that FPD has a role to play in speech perception is not specific to exemplar models and is also found in at least some abstractionist models. Exemplar models do diverge from abstractionist models, however, in assuming that in addition to being relevant to on-line speech perception and understanding, FPD is stored in long-term memory. More specifically, and as opposed to abstractionist models, exemplar models posit that lexical representations are phonetically rich.

To our knowledge, much of the available evidence for long-term storage of FPD in the mental lexicon comes from production studies. As shown by Bybee (2001, 2006a) and Pierrehumbert (2001, 2002), frequency-dependent differences in the phonetic realization of words that meet the same structural description (e.g., words underlyingly containing a schwa followed by a sonorant, such as the high-frequency word *every* [evɹi], produced with no schwa, compared with the mid-frequency word *memory* [memɻi], produced with a syllabic /r/), must be learned and stored in the mental lexicon by speakers in the course of language acquisition. This is also true of sociophonetic variation, which has to be learned inasmuch as the relationship between phonetic forms and social categories is arbitrary (Foulkes and Docherty, 2006). Because these sometimes subtle patterns of phonetic variation have to be detected by the speaker before she/he proves able to reproduce them, these studies lend strong albeit indirect support to the assumption that perceived FPD is stored in the lexicon. More direct evidence for this assumption is also available from a variety of sources. In a well-known series of experiments, Goldinger (1996, 1998) showed that prior exposure to a speaker's voice facilitates later recognition of words spoken by the same speaker as opposed to a different speaker. Strand (2000) found that listeners respond more slowly to nonstereotypical male and female voices than to stereotypical voices in a speeded naming task. These studies suggest that the individual characteristics of the speaker's voice as well as the acoustic/phonetic properties associated with the speaker's gender are retained in memory by listeners. Both Johnson (1997b) and Goldinger (1998) consider that direct storage of FPD in the

lexicon allows listeners to deal with between-speaker variations in the production of words without having to resort to a normalization procedure (but see Mitterer, 2006, for experimental counterevidence).

Little is known yet about the shape that exemplars stored in memory may have. A survey of the relevant literature indicates that exemplars are generally considered as multimodal sensory-motor representations of language chunks of various size (more on this later), and can therefore be characterized in a general way as being *a)* non-symbolic, *b)* parametric, *c)* in a relationship of intrinsic similarity with the input speech signal, and *d)* abstract, up to a certain extent, given the limits inherent in the auditory system, such as the fact that the auditory trace of speech has been shown to fade away after 400 ms (Pardo and Remez, 2007). Because of the limits of our current knowledge in that domain, current exemplar models of speech production and perception can be, in a paradoxical way, far more abstract than they purport to be. In these models, exemplars are sometimes represented in a highly schematized form which bears little resemblance with the fine-grained acoustic structure of speech. Much research still needs to be done to better characterize what exemplars may look like in the listener's brain.

Proposals made by researchers within the exemplar framework do not form a fully uniform set and there are important differences between them. For example, Hintzman's (1986) rationalist approach to memory may be opposed to the neo-empiricist viewpoint advocated by Coleman (2002). Importantly, there is a lack of consensus among proponents of the exemplar approach as regards the status of phonological representations in speech perception. In some models, such as Johnson's (1997a; 2005a) XMOD model, exemplars have no internal structure, and are conceived as unanalyzed auditory representations associated with whole words. This, however, does not mean that sublexical units such as segments and syllabic constituents cannot have a psychological reality. Although it is assumed that such units are not explicitly represented in memory, they can nevertheless be brought to the listener's consciousness as the speech signal is being mapped onto the lexicon. These units temporarily emerge as a by-product of lexical activation, as connections between time-aligned, phonetically-similar portions of exemplars are established. In this framework, listeners are assumed to be simultaneously sensitive to units of different sizes in the speech signal, albeit with a natural bias for larger units to prevail over smaller ones (Grossberg, 2003). What may be viewed as a phonological structure, with a certain degree of abstraction, is therefore built up by the listener in the online processing of speech, al-

though this structure is said to be but "a fleeting phenomenon - emerging and disappearing as words are recognized" (Johnson, 1997a). Other researchers (Hawkins, 2003, 2007; Luce and McLennan, 2005; McLennan and Luce, 2005; Pierrehumbert, 2006) have proposed a hybrid approach in which exemplars are encoded in memory in conjunction with permanently-stored abstract phonological representations. In Hawkins' POLYSP model of speech perception and understanding (Hawkins and Smith, 2001; Hawkins, 2003, 2007) for example, fine phonetic detail is mapped onto abstract prosodic structures as characterized in the Firthian Prosodic Analysis phonological framework.

The whole-word exemplar hypothesis, as it is adopted in some models of speech perception, raises a number of issues that have been highlighted by different authors. First, it is not always clear why words should indeed be postulated as basic units of processing and storage, rather than fragments of speech of many different sizes as these would empirically come to surface in the utterances to which listeners are exposed in the course of their life. If the logic that governs the exemplar approach is to be fully followed, one should assume that sequences of words of high frequency, such as *I don't know*, should be stored as single units in what then becomes a highly extended mental lexicon (see Bybee, 2001, 2006b). Second, the whole-word exemplar hypothesis in perception is inconsistent with what Pierrehumbert (2006) refers to as the *phonological principle*, i.e., "that languages have basic building blocks, which are not meaningful in themselves, but which combine in different ways to make meaningful forms", as shown by the fact that classically-defined allophonic rules are found to apply to a large majority of words sharing the same structural description, even if they may not extend to all of these words. Pierrehumbert (2006) also points out that it is difficult to see how whole-word exemplar models can account for the *bistable* character of speech perception, i.e., that an ambiguous speech sound potentially associated with two categories will be perceived as a member of one and only one of these categories at any one time (since such response patterns seem to rely on a winner-take-all competition among two underlying abstract units).

There is a well-known and extensive body of evidence in favor of the assumption that infra-lexical phonological representations come into play in spoken word recognition (e.g., Cutler et al., 2007; Lahiri and Marslen-Wilson, 1991; Lahiri and Reetz, 2002; Pallier, 2000). In addition, numerous experimental studies have shown that, in some circumstances at least, abstract phonological categories seem to prevail upon fine phonetic detail in the mapping of speech sounds onto meaning, as demonstrated by Lahiri and Reetz

(2002), among others. In the following section, we focus on two studies that our coworkers and we recently carried out and whose results also point to the role of abstract phonological representations in speech perception.

3. Further evidence for the role of abstract phonological representations in speech perception

Dufour, Nguyen and Frauenfelder (2007) examined the influence that regional differences in the phonemic inventory of French may have on how spoken words are recognized. Whereas the phonemic system of standard French is traditionally characterized as containing three mid vowel pairs, namely /e/-/ɛ/, /ø/-/œ/, and /o/-/ɔ/, as in *épée* /epe/ "sword" vs. *épais* /epɛ/ "thick", and *côte* /kot/ "hill" vs. *cote* /kɔt/ "rating", southern French is viewed as having three mid-high vowel phonemes only, /e/, /ø/ and /o/ (Durand, 1990). [ɛ], [œ] and [ɔ] appear at the phonetic level but they are in complementary distribution with respect to the corresponding mid-high variants, according to a variant of the so-called *loi de position* (a mid-vowel phoneme is realized as mid-high in an open syllable and as mid-low in closed syllables and whenever the next syllable contains schwa, see Durand, 1990). Thus, *épée* and *épais* will be both pronounced [epe] and *côte* and *cote* will be both pronounced [kɔtə] in southern French. Dufour et al. (2007) asked how words such as *épée*, *épais*, *côte* and *cote*, as produced by a speaker of standard French, i.e., with a contrast in vowel height in the word-final syllable, were perceived by speakers of both standard and southern French. Using a lexical decision task combined with a long-lag repetition priming paradigm, Dufour et al. found that pairs of words ending in a front mid vowel (e.g. *épée - épais*) were not processed in the same way by both groups of subjects. Standard French speakers perceived the two words as being different from each other, as expected, whereas southern French speakers treated one word as a repetition of the other. By contrast, both groups of subjects perceived the two members of /o/-/ɔ/ word pairs as different one from the other. Thus, the results showed that there are within-language differences in how isolated words are processed, depending on the listener's regional accent. Note that southern speakers are far from being unfamiliar with standard French. On the contrary, they are widely exposed to it through the media and at school in particular. According to Dufour et al., the observed response patterns for southern speakers may be accounted for by assuming that the /o/-/ɔ/ contrast is better defined than the /e/-/ɛ/ contrast in these speakers' receptive phonological

knowledge of standard French. The /o/-/ɔ/ contrast is a well-established and highly recognizable feature of standard French, which is as such well-known to southern speakers, even if this contrast is neutralized in these speakers' dialect. By comparison, the distribution of /e/ and /ɛ/ in word-final position in standard French is characterized by greater complexity both across and within speakers, and there is evidence showing that word-final /e/ and /ɛ/ are in the process of merging in Parisian French (although the speaker used in Dufour et al.'s study did make the distinction between the two vowels, as confirmed by the fact that standard French subjects did not process the two carrier words as being the second one a repetition from the first one). Dufour et al. hypothesized that because of the unstable status of the /e/-/ɛ/ contrast, both vowels were perceptually assimilated to the same abstract phonological category by speakers of southern French.

Nguyen, Wauquier-Gravelines, Lancia and Tuller (2007b) recently undertook a study on the perceptual processing of liaison consonants in French. Liaison in French is a well-known phenomenon of external sandhi that refers to the appearance of a consonant at the juncture of two words, when the second word begins with a vowel, e.g. *un* [œ̃] + *enfant* [ɑ̃fɑ̃] → [œ̃nɑ̃fɑ̃] "a child", *petit* [pəti] + *ami* [ami] → [pətitami] "little friend". In earlier work, Wauquier-Gravelines (1996) showed that listeners found it more difficult to detect a target phoneme (e.g., /n/) in a carrier phrase, when that phoneme was a liaison consonant (*son avion* [sɔ̃navjɔ̃] "her plane") compared with a word-initial consonant (*son navire* [sɔ̃naviʁ] "her ship"). The proportion of correct detection proved significantly lower for the liaison than for the word-initial target consonant. According to Wauquier-Gravelines, the listeners' response pattern was attributable to the specific phonological status that liaison consonants have in French. More particularly, and in the autosegmental phonology framework (Encrevé, 1988) espoused by Wauquier-Gravelines, liaison consonants are treated as floating segments with respect to both the segmental and syllabic tiers, as opposed to fixed segments, which are are lexically anchored to a skeletal slot, and which include word-initial consonants, but also word-final (e.g., *la bonne* [labɔn] "the maid") and word-internal (e.g., *le sénat* [ləsena] "the senate") ones. Using a speeded phoneme detection task, Nguyen and colleagues (2007b) aimed to confirm that detecting liaison consonants in speech is difficult. They examined to what extent differences in the detection rate of liaison consonants vs. word-initial consonants could, at least in part, stem from the phonetic properties of these consonants, by systematically manipulating these properties. In addition, the potentially distinctive status of

liaison consonants compared with fixed consonants in perception was further explored by inserting word-final and word-medial fixed consonants as well as word-initial ones in the material. The results revealed that the target consonant's fine phonetic properties had no measurable influence on percent correct detection. They also showed that listeners tended to miss liaison consonants more often than fixed consonants, whether these were in word-initial, word-final or word-medial position. Nguyen and colleagues (2007b) pointed out that the listeners' response pattern was partly consistent with an exemplar-based theory of French liaison such as the one proposed by Bybee (2001). In this approach, liaison consonants are deeply entrenched in specific grammatical constructions, and the realization of liaison is highly conditioned by the strength of the associations between words within such constructions. Although little is said in Bybee's theory about how liaison consonants may be processed in speech understanding, a prediction that may be derived from this theory is that listeners will process liaison consonants as part and parcel of the constructions in which these consonants appear. As a result, it may be difficult for listeners to identify liaison consonants as context-independent phonemic units, as explicitly required in a phoneme-detection task. This, however, should be true for *all* the segments a construction may contain. On this account, listeners should not experience less difficulty in detecting a word-initial consonant compared with a liaison consonant, when these consonants appear in word sequences that are highly similar to each other with respect to their morpho-syntactic and phonetic make-up, as was the case in Nguyen and colleagues' (2007b) material. The lower detection rates observed for liaison than for word-initial target consonants was consistent with the assumption that liaison consonants have a specific phonological status and, to that extent, provided better evidence for the abstractionist autosegmental account of liaison than for the exemplar-based account. Readers are referred to Nguyen and colleagues (2007b) for further detail about the experiment and its potential theoretical implications.

4. The dynamical view to speech perception: Beyond the exemplars vs. abstractions dichotomy?

The above review of the literature suggests that the dichotomy that is sometimes established between the exemplar-based and abstractionist approaches to speech perception is to a large extent artificial. Experimental evidence is available that provides support for the role of both fine phonetic detail and

abstract phonological categories in speech perception. The recent development of so-called hybrid models (Hawkins, 2003, 2007; Luce and McLennan, 2005; McLennan and Luce, 2005; Pierrehumbert, 2006) is governed by the assumption that FPD and abstract phonological categories are combined with each other in the representations associated with words in the speaker/listener's memory.

Over the last decade, Tuller and colleagues (Case, Tuller and Kelso, 1995; Tuller, Case, Ding and Kelso, 1994; Tuller, 2003, 2004) have developed a model that shares some of the characteristics of the hybrid approach. This model, referred to as the TCDK model hereafter, uses concepts from the theory of nonlinear dynamical systems to account for the mechanisms involved in the categorization of speech sounds. In this model, there are two complementary aspects to speech perception. On the one hand, speech perception is assumed to be a highly context-dependent process sensitive to the detailed acoustic structure of the speech input. On the other hand, it is viewed as bringing into play a non-linear dynamical system characterized by a limited number of stable states, or attractors, which allow the system to perform a discretization of the perceptual space, and which are associated with abstract perceptual categories. In this section, and after a brief and schematic presentation of the TCDK model, we report the results of a study recently conducted on the categorization of speech sounds in French with a view to testing some of the model's predictions. The implications of the model for the exemplar vs. abstraction debate will then be discussed.

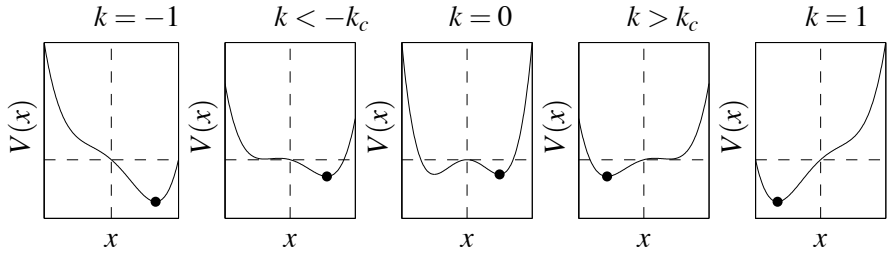
The model was first designed to account for listeners' response patterns in a binary-choice speech categorization task. In the experiments reported in Tuller et al. (1994), listeners were presented with stimuli ranging on an acoustic continuum between *say* and *stay* and their task was to identify each stimulus as either of these two words. Listeners' responses were modeled using a non-linear dynamical system governed by the following equation:

$$V(x) = kx - \frac{x^2}{2} + \frac{x^4}{4}$$

In this equation, x represents a one-dimensional perceptual form, k a control parameter, and $V(x)$ a potential function which may have up to two minima, depending on the value of k. The control parameter k is itself a function of the acoustic characteristics of the stimulus, on the one hand, and the combined effects of learning, linguistic experience and attentional factors, on the other hand. For a given value of k, the system's state evolves in the x percep-

tual space to get trapped into a local minimum, or attractor, of $V(x)$. Each of the two possible listener's responses in the categorization task corresponds to one attractor in the perceptual space. Figure 1 shows the shape of the potential function for five values of k comprised between -1 and 1. The potential function has one minimum only for extreme values of k, which correspond to stimuli unambiguously associated with either of the two categories, and two minima in the middle range of k, where both categories are possible. As k increases in a monotonic fashion (from left to right in Figure 1), and in the vicinity of a critical value k_c, the system's state, represented by the filled circle in Figure 1, abruptly switches from the basin of attraction in which it was initially located, to the second basin that has gradually formed as the first one disappears.

Figure 1. Shape of the potential function $V(x)$ for five values of k. Adapted from Tuller et al. (1994).



In Tuller and colleagues' (1994) experiments, the stimuli on the *say-stay* continuum were presented to the listener in either a randomized order, or a sequential order. In the latter case, listeners heard the entire set of stimuli twice, going from one of the two endpoints (e.g., *say*) to the other (*stay*), and then back to the first one (*say*) again. In such sequential presentations, three possible response patterns were to be expected: *a)* hysteresis, defined as the tendency for the listener's response at one endpoint to persist across the ordered sequence of stimuli towards the other endpoint, *b)* enhanced contrast, which, as opposed to hysteresis, refers the case where listener quickly switches to the alternate percept and does not hold on to initial categorization, *c)* critical boundary, where the switch between the two percepts is associated with the same stimulus regardless of the direction of presentation across the continuum. The results showed that critical boundary was much less frequent than hysteresis and contrast, which occurred equally often. These data provided strong support for the assumption that speech perception is a highly context-

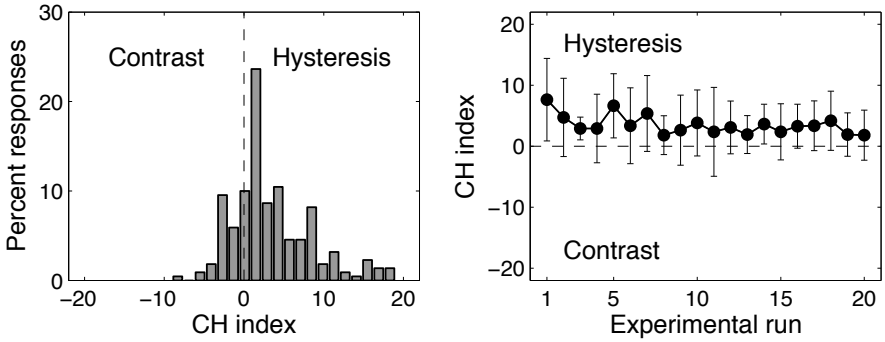
dependent process, characterized by a rich variety of dynamical properties. Readers are referred to Tuller et al. (1994) and Case et al. (1995) for further detail on these experiments.

Nguyen, Lancia, Bergounioux, Wauquier-Gravelines and Tuller (2005) and Nguyen, Lancia and Tuller (2007a) recently undertook to extend Tuller and colleagues' (1994) hypotheses and experimental paradigm to the categorization of speech sounds in French. In Nguyen, Lancia and Tuller (2007a), the material was made up of 21 stimuli on an acoustic continuum between *cèpe* /sɛp/ "mushroom" and *steppe* /stɛp/ "steppe". Each stimulus contained a silent interval between /s/ and /ɛ/ whose duration increased from 0 ms (Stimulus 1) to 100 ms (Stimulus 21) in 5-ms steps. These stimuli were used in a speeded forced-choice identification task administered to eleven native speakers of French naive as to the purposes of the experiment and with no known hearing defects. Listeners were presented with the 21 stimuli in both random and sequential (1-21-1 or 21-1-21) order. The experiment comprised 20 randomized presentations alternating with 20 sequential presentations. The inter-stimuli interval was two seconds and the experiment lasted about an hour.

An index referred to as the CH (Contrast-Hysteresis) index was devised to measure the amount of hysteresis or, on the opposite, enhanced contrast, in each subject's responses to each sequential presentation. This entailed locating the position on the continuum of the stimulus associated with the switch from one response to the other in the first part of the presentation, on the one hand, and in the second part of the presentation, on the other hand. The distance between these two points was then measured, in a such a way that positive values corresponded to hysteresis, negative values to enhanced contrast, and 0 to critical boundary. The distribution of the CH index across the 20 sequential presentations for all the subjects is shown in Figure 2. These data indicate that hysteresis prevailed upon enhanced contrast and critical boundary. The CH index reached a grand average value of 3.5 that proved to be significantly higher than 0 in a linear mixed-effects model using the CH values as the predicted variable, the intercept as predictor (fixed effect) and the subjects as blocking factor (random effect; $t(208) = 3.93, p < 0.001$).

Figure 2 also shows the mean and standard deviation of the CH index for each of the 20 sequential presentations, in chronological order from the beginning of the experiment. An important prediction of the TCDK model is that the amount of hysteresis should decrease as the subject gets more experienced with the task and stimuli (e.g., Tuller, 2004; Nguyen, Lancia, Bergounioux,

Figure 2. Observed values of the CH index in Nguyen, Lancia and Tuller's (2007a) speech categorization experiment. Left panel: distribution across all the presentations; right panel: mean value and standard deviation for each of the 20 sequential presentations.



Wauquier-Gravelines & Tuller, 2005). This prediction was borne out by the data, as a decrease in the CH value over the course of the experiment was observed which proved statistically significant in a linear mixed-effects model using the CH value as predicted variable, the rank of presentation as predictor and the subjects as blocking factor ($t(208) = -2.792, p < 0.01$). These results offered further confirmation that the speech perception system can be modelled as a nonlinear dynamical system whose current state simultaneously depends on the input speech sound, the system's past state, and higher-level cognitive factors that include the listener's previous experience with the sounds she/he has to categorize. More detail about the experimental design and results is available in Nguyen et al. (2005, 2007a). In the following, we concentrate on how the TCDK model may contribute to addressing the general issues that have been discussed in the present paper.

In spite of its being linked to a specific experimental task (forced-choice categorization), the TCDK model shows a number of general properties that, in our view, open the way towards a novel, hybrid view to speech perception that gets beyond the dichotomy traditionally established between exemplars and abstract phonological representations. Clear differences arise between the non-linear dynamical approach exemplified by the TCDK model and the abstractionist approach. In the former, to a greater extent and more systematically than in the latter, speech categorization is viewed as being sensitive to the detailed acoustic characteristics of the input signal. For example, it

is assumed that small variations in the acoustic structure of an ambiguous speech sound, that may correspond to two different categories, can lead to large changes in the perceptual system's response. This will be the case when these variations cause the system to move across a saddle point in the potential function (see Figure 1). Note, however, that perceptual sensitivity to small acoustic change is not assumed to be the same in all regions of the acoustic space. Variations shown by a stimulus close to a prototypical sound unambiguously associated with a given perceptual category, will have little impact on the listener's response, which will then be viewed in the model as being governed by a one-attractor potential function (see left and right panels of Figure 1). The TCDK model also predicts that the relative stability of a category will vary depending on how frequently that category has been perceived in the preceding sequence of speech sounds. This is attributed to the fact that the tilt of the potential function changes more sharply in response to a variation in the input sound, when a given category has been perceived more often. Yet another prediction of the model is that the location of the perceptual switch from one category to another in the acoustic space tends strongly to depend on the trajectory followed by the stimuli in that space, as is the case in both hysteresis and enhanced contrast. Over a longer time scale, increasing experience with the stimuli is expected to affect the dynamics of speech categorization, which will tend to move away from hysteresis towards enhanced contrast. The predicted sensitivity of the speech perception system to gradient acoustic properties, frequency of occurrence of perceived categories, trajectory of speech sounds in the acoustic space, and training, is consistent with listeners' observed response patterns, and seems difficult to account for by abstractionist models of speech perception. In the TCDK model, this sensitivity partly derives from the fact that speech sounds are mapped onto a discrete and finite set of perceptual categories by means of a *continuous* potential function, as opposed to the sharp division between sounds and percepts often posited in abstractionist models.

While the dynamical nature of speech categorization is central to the TCDK model, it is also, and up to a certain extent, emphasized in the exemplar approach. Exemplars are assumed to accumulate in memory as listeners are exposed to them, and this causes boundaries between categories to be continuously pushed around in the perceptual space. As a result, more frequent categories (represented by a higher number of exemplars) gradually come to perceptually prevail upon less frequent ones (Pierrehumbert, 2006). Perceptual categories are taken in the exemplar approach to be time-dependent,

and to continuously evolve in the course of the conversational interactions in which speakers/listeners engage themselves. This of course has major theoretical implications, as frequency of use is expected to have an impact on the very form of phonological representations in memory. As indicated above, however, dynamics in speech perception is not restricted to the incremental effect of exposure on perceived categories, and encompasses a much wider range of phenomena such as hysteresis, contrast, bifurcation, and stability. The TCDK model aims to take advantage of the powerful theory of non-linear dynamical systems to account for these phenomena in all their variety and along short as well as long time scales. It offers an explanation of the bistability of speech perception, attributed to the coexistence of two mutually exclusive attractor states in the perceptual space. In addition, theoretical and methodological tools (e.g., Erlhagen et al., 2006) are available that may allow the non-linear dynamical framework to be extended to the study of conversation interaction between two or several speakers, and to model the dynamics of speech processing and its influence on the organization of perceived categories as this interaction unfolds in time.

5. Acknowledgements

This work was partly supported by the ACI Systèmes complexes en SHS Research Program (CNRS & French Ministry of Research). A first version of the present paper was presented at the Workshop on Phonological Systems and Complex Adaptive Systems held in Lyons in July 2005. Feedback from Abby Cohn, Adamantios Gafos, and Sharon Peperkamp, among other participants, is gratefully acknowledged. Revised versions were presented later in seminars held at the Université de Toulouse - Le Mirail (ERSS laboratory) and the Université de Paris X (MoDyCo laboratory). We thank Joaquim Brandão de Carvalho, Jacques Durand, John Goldsmith, and Bernard Laks, in particular, for critical comments and suggestions.

Bibliography

Allen, J. and Miller, J. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 115:3171–3183.

Andruski, J., Blumstein, S., and Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, 52:163–187.

Barlow, M. and Kemmer, S., editors (2000). *Usage-Based Models of Language*. Center for the Study of Language and Information, Stanford, CA.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge, UK.

Bybee, J. (2006a). *Frequency of Use and the Organization of Language*. Oxford University Press, Oxford, UK.

Bybee, J. (2006b). From usage to grammar: the mind's response to repetition. *Language*, 82:529–551.

Bybee, J. and McClelland, J. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, 22:381–410.

Case, P., Tuller, B., Ding, M., and Kelso, J. (1995). Evaluation of a dynamical model of speech perception. *Perception and Psychophysics*, 57:977–988.

Case, P., Tuller, B., and Kelso, J. (2003). The dynamics of learning to hear new speech sounds. *Speech Pathology*.

Coleman, J. (2002). Phonetic representations in the mental lexicon. In Durand, J. and Laks, B., editors, *Phonetics, Phonology, and Cognition*. Oxford University Press, Oxford, UK.

Couper-Kuhlen, E. and Ford, C., editors (2004). *Sound Patterns in Interaction. Cross-linguistic Studies from Conversation*. John Benjamins, Amsterdam, The Netherlands.

Cutler, A., Eisner, F., McQueen, J., and Norris, D. (2007). Coping with speaker-related variation via abstract phonemic categories. In Fougeron, C., D'Imperio, M., Kühnert, B., and Vallée, N., editors, *Papers in Laboratory Phonology X*. Mouton de Gruyter, Berlin, Germany. to appear.

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, 28:441–465.

Dufour, S., Nguyen, N., and Frauenfelder, U. (2007). The perception of phonemic contrasts in a non-native dialect. *Journal of the Acoustical Society of America Express Letters*. accepted for publication.

Durand, J. (1990). *Generative and Non-Linear Phonology*. Longman, Londres.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. and McClelland, J. (1988). Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27:143–165.

Encrevé, P. (1988). *La liaison avec et sans enchaînement*. Seuil, Paris.

Erlhagen, W., Mukovskiy, A., and Bicho, E. (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Research*, 1083:174–188.

Eulitz, C. and Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience*, 16:577–583.

Fitzpatrick, J. and Wheeldon, L. (2000). Phonology and phonetics in psycholinguistics models of speech perception. In Burton-Roberts, N., Carr, P., and Docherty, G., editors, *Phonological Knowledge: Conceptual and Empirical Issues*, pages 131–160. Oxford University Press, Oxford, UK.

Foulkes, P. and Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34:409–438.

Gaskell, M. (2003). Modelling regressive and progressive effects of assimilation in speech perception. *Journal of Phonetics*, 31:447–463.

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, N., and Coupland, J., editors, *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pages 1–68. Cambridge University Press, Cambridge, UK.

Goldinger, S. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 22:1166–1183.

Goldinger, S. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105:251–279.

Goldstein, L. (2003). Emergence of discrete gestures. In *Proceedings of the XVth International Congress of Phonetic Sciences*, pages 85–88, Barcelona, Spain.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31. 423–445.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31:373–405.

Hawkins, S. (2007). Phonetic variation as communicative system: Perception of the particular and the abstract. In Fougeron, C., D'Imperio, M., Kühnert, B., and Vallée, N., editors, *Papers in Laboratory Phonology X*. Mouton de Gruyter, Berlin. to appear.

Hawkins, S. and Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Rivista di Linguistica*, 13:99–188.

Hintzman, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93:411–428.

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics*, 50:101–113.

Johnson, K. (1997b). Speech perception without speaker normalization. In Johnson, K. and Mullenix, J., editors, *Talker Variability in Speech Processing*, pages 145–166. Academic Press.

Johnson, K. (2005a). Decisions and mechanisms in exemplar-based phonology. *UC Berkeley Phonology Lab Annual Report*, pages 289–311.

Johnson, K. (2005b). Speaker normalization in speech perception. In Pisoni, D. and Remez, R., editors, *The Handbook of Speech Perception*. Blackwell, Malden, MA. 363–389.

Klatt, D. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7:279–312.

Lahiri, A. and Marslen-Wilson, W. (1991). The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition*, 38:245–294.

Lahiri, A. and Reetz, H. (2002). Underspecified recognition. In Gussenhoven, C. and Warner, N., editors, *Papers in Laboratory Phonology VII*, pages 637–675. Mouton de Gruyter, Berlin, Germany.

Luce, P. and McLennan, C. (2005). Spoken word recognition: The challenge of variation. In Pisoni, D. and Remez, R., editors, *The Handbook of Speech Perception*, pages 591–609. Blackwell, Malden, MA.

Marslen-Wilson, W. and Warren, P. (1994). Levels of perceptual representation and process in lexical access - words, phonemes, and features. *Psychological Review*, 101:653–675.

Martin, J. and Bunnell, H. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, 69:559–567.

Martin, J. and Bunnell, H. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 8:473–488.

McClelland, J. and Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18:1–86.

McLennan, C. and Luce, P. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, 31:306–321.

McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86:B33–B42.

McMurray, B., Tanenhaus, M., Aslin, R., and Spivey, M. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category vot on lexical access. *Journal of Psycholinguistic Research*, 32:77–97.

Meltzoff, A. and Moore, M. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting*, 6:179–192.

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? *Phonetica*, 63:209–229.

Nguyen, N., Lancia, L., Bergounioux, M., Wauquier-Gravelines, S., and Tuller, B. (2005). Role of training and short-term context effects in the identification of /s/ and /st/ in French. In Hazan, V. and Iverson, P., editors, *ISCA Workshop on Plasticity in Speech Perception (PSP2005)*, pages A38–39, London, UK.

Nguyen, N., Lancia, L., and Tuller, B. (2007a). The dynamics of speech categorization: Evidence from French. in preparation.

Nguyen, N., Wauquier-Gravelines, S., Lancia, L., and Tuller, B. (2007b). Detection of liaison consonants in speech processing in French: Experimental data and theoretical implications. In Prieto, P. and Solé, M., editors, *Laboratory Approaches to Romance Phonology*. John Benjamins. in press.

Pallier, C. (2000). Word recognition: do we need phonological representations? In Cutler, A., McQueen, J., and Zondervan, R., editors, *Proceedings of the Workshop on Spoken Word Access Processes (SWAP)*, pages 159–162, Nijmegen.

Pallier, C., Colomé, A., and Sebastián-Gallès, N. (2001). The influence of native-language phonology on lexical access: exemplar-based vs. abstract lexical entries. *Psychological Science*, 12:445–449.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119:2382–2393.

Pardo, J. S. and Remez, R. E. (2007). The perception of speech. In Traxler, M. and Gernsbacher, M., editors, *The Handbook of Psycholinguistics, Second Edition*. Elsevier, Cambridge, MA. in press.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. and Hopper, P., editors, *Frequency effects and the emergence of linguistic structure*, pages 137–157. John Benjamins, Amsterdam.

Pierrehumbert, J. (2002). Word-specific phonetics. In Gussenhoven, C. and Warner, N., editors, *Papers in Laboratory Phonology VII*, pages 101–140. Mouton de Gruyter, Berlin, Germany.

Pierrehumbert, J. (2006). The next toolkit. *Journal of Phonetics*, 34:516–530.

Pitt, M. and Johnson, K. (2003). Using pronunciation data as a starting point in modeling word recognition. In *Proceedings of the XVth International Congress of Phonetic Sciences*, Barcelona, Spain.

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111:1872–1891.

Stevens, K. (2004). Invariance and variability in speech: Interpreting acoustic evidence. In Slifka, J., Manuel, S., and Matthies, M., editors, *Proceedings of From Sound to Sense: 50+ Years of Discovery in Speech Communication*, pages B77–B85, Cambridge, MA. MIT. URL: `www.rle.mit.edu/soundtosense/`.

Strand, E. (2000). *Gender Stereotype Effects in Speech Processing*. PhD thesis, Ohio State University.

Studdert-Kennedy, M. (2002). Mirror neurons, vocal imitation and the evolution of particulate speech. In Stamenov, M. and Gallese, V., editors, *Mirror Neurons and the Evolution of Brain and Language*, pages 207–227. John Benjamins, Amsterdam.

Tuller, B. (2003). Computational models in speech perception. *Journal of Phonetics*, 31:503–507.

Tuller, B. (2004). Categorization and learning in speech perception as dynamical processes. In M.A., R. and Van Orden, G., editors, *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*. National Science Foundation. URL: `www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp`.

Tuller, B., Case, P., Ding, M., and Kelso, J. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 20:3–16.

Vihman, M. (2002). The role of mirror neurons in the ontogeny of speech. In Stamenov, M. and Gallese, V., editors, *Mirror Neurons and the Evolution of Brain and Language*, pages 305–314. John Benjamins, Amsterdam.

Wauquier-Gravelines, S. (1996). *Organisation phonologique et traitement de la parole continue*. Unpublished phd dissertation, Université Paris 7, Paris.

West, P. (1999). Perception of distributed coarticulatory properties in English /l/ and /ɹ/. *Journal of Phonetics*, 27:405–426.