Temporal integration in the perception of speech: Introduction

Noël Nguyen, Sarah Hawkins

Temporal integration in the perception of speech: Introduction

Noël Nguyen (1) and Sarah Hawkins (2)

(1) Laboratoire Parole et Langage, CNRS & Université de Provence, 29 avenue Robert Schuman, 13621 Aix-en-Provence, France

(2) Department of Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, United Kingdom

September 29, 2003

running title: Temporal integration in the perception of speech

1 Introduction

In studies of speech perception, temporal integration refers to how chunks of information arriving at the ears at different times are linked together by the listener in mapping speech sounds onto meaning. Classical models focused on the perceptual grouping of acoustic cues contained in short stretches of time in the identification of phonetic segments. In recent years, however, a different view has emerged as speech perception has come to be studied within a broader context and from a multidisciplinary perspective. Thus, the relevance of non-local, long-domain cues to phonological contrasts has been demonstrated. The status of the phonetic segment as a basic perceptual unit has been debated. And the primacy of the auditory channel over the visual channel has been questioned. These issues have profound implications for how temporal integration is defined and accounted for.

Temporal integration in the perception of speech was the unifying focus of the TIPS international workshop held in Aix-en-Provence in April 2002. The workshop comprised fifteen focus papers along with commentaries by as many discussants. Written versions of most of the focus papers and commentaries are now offered in this special issue of the Journal of Phonetics[1]. Central to the workshop was how the mind combines information in the speech signal to understand the speaker's message, when information about a particular phonological or grammatical distinction may occur in matters of milliseconds or over several syllables. Another major goal of the workshop was to engage in dialogue with researchers from outside phonetics. Interfaces between current phonetic approaches to speech understanding and related but distinct disciplines are especially valuable in helping this interdisciplinary field to move ahead effectively. Thus this volume includes papers on subjects that would ordinarily be considered as falling at the edge of, or even outside, the scope of the Journal of Phonetics. These subjects include psycholinguistics, psychoacoustics, neuropsychology and computational modeling. We believe that these papers enrich the current volume by allowing a more comprehensive exploration of the many facets of temporal integration by the human brain. Papers in this volume that fall into these categories include those by Grossberg,

Gaskell, Mody, Rosen, de Cheveigné, and Moore, together with their associated commentaries, especially those by King, Scott, Gow, Shamma, Tuller, Cooke, Macar, and Kello. By their nature, these papers contain a significant amount of tutorial material rather than original research, but all offer insight of value to phoneticians and others seeking to place phonetic processes within the broader field of auditory processing. Moreoever, most of these authors have taken account of the particular challenges to his or her own area represented by some of the less-well-known phonetic observations found in, for example, Local's and Coleman's papers. We thus hope that the benefits from this cross-disciplinary exchange go in both directions.

The volume is subdivided into six main sections, reflecting the structure of the conference itself. Each section comprises from one to three focus papers followed by two or three commentaries. The initial keynote section sets the scene, while each of the following five sections is organised around a common theme: phonetic and phonological issues, computational modelling, pathology (dyslexia and specific language impairment), psychoacoustics, and development. Most papers were circulated in a preliminary version among the authors prior to TIPS and were of course extensively discussed during the workshop itself and in the preparation of this volume. Because of the many exchanges that took place between the authors, and of the intrinsic complementarity of the focus papers and the associated commentaries, this volume comprises a coordinated set of papers that should broaden perspectives and further understanding of the mechanisms involved in the perception of speech.

Temporal integration is a complex problem that may be approached from many different angles. This is partly because time itself is a ubiquitous concept. As pointed out by several authors in this volume, there are many different time scales that are relevant to speech communication, extending from the very short (acoustic cues contained in a stop burst), to the long (phonetic exponents of syllable or word prosodies), and the very long (phonetic patterns that signal the end of a conversational turn, for example). Time is also intrinsic to the speech processing system itself, whose properties evolve in a continuous manner. In addition to the well-known modifications during evolution and ontogeny, recent research on exemplar memory suggests that the speech

processing system remains highly plastic in adults, and continuously adapts itself to the detailed phonetic characteristics of the speech input. The complex nature of temporal integration is yet more apparent as one moves away from laboratory speech and into the domain of conversational interaction. We then need to understand how the listener succeeds in differentiating the speech signal originating from a single talker from other streams of sensory information. As the talkers interact with each other and engage in an entrainment process, we also need to consider the fact that temporal integration may no longer be a process occurring in the brain of one person passively listening to another, but may rather be conditioned by the behaviour of both participants.

2 Outline of the volume

2.1 Setting the scene

The volume begins with three focus papers (Remez, Local, Goldinger & Azuma) and two commentaries (Docherty, Nygaard) that set the stage, identify the basic issues TIPS is focussed upon, and make several key statements. One theme is that, for the representations and processes employed by the listener to be adequately characterized, speech perception has to be studied within its natural site of occurrence, that is communication in noise. Another theme is that, rather than there being one basic unit of speech perception, there may instead be a variety of competing candidates whose temporal domain depends not only on phonological, lexical and grammatical factors, but also on the dynamics of conversational interaction and on the particular demands of the experimental or other listening situation in which the listeners are placed. While the possibility of more than one basic unit of perception has long been recognized, placing putative units in the wider interactional context is not only unusual in acoustic phonetics, but highlights the limits of current understanding and the diversity of the factors that could influence speech processing (though cf. Lindblom's (1990) H&H theory).

2.2 Phonetic and phonological issues

In traditional models of speech perception, the speech signal is converted by the listener into a sequence of abstract segmental units which are in turn mapped onto the lexicon. In this view, the role of phonetic detail is confined to the early stages of processing, and does not extend beyond the setting up of a segmental representation of the utterance. Recent work, however, suggests that phonetic detail may perform a central role in the understanding of meaning. In exemplar-based models of speech perception, words and frequently-used grammatical constructions are associated in memory with multimodal, detailed exemplars which allow fine phonetic information to be directly related to meaning. The two focus papers (Coleman, Hawkins) and one of the three commentaries (Laver) in this section are concerned with the form and function of phonetic and phonological representations in an exemplar-based approach to the perception and understanding of speech. The second commentary (King) focusses on the implications of phonetic knowledge for automatic speech recognition and speech synthesis, and the third commentary (Scott) discusses neuropsychological evidence for hierarchical structure in language representation. One question is how phonological representations may come to emerge from exemplars. Another issue is to what extent phonological representations, and more generally formal linguistic units of analysis, are involved in the access to meaning.

2.3 Computational modelling

A variety of mechanisms have been invoked to explain how temporal integration is performed by the listener. Traditional computational models of speech perception tended to resort to a spatial encoding of time, e.g. by representing a word as an array of features that was presented to the model as a bidimensional image and in one pass. This unsatisfactory solution has been abandoned as new models have emerged that have the capacity to process signals that unfold in time. Another major improvement is that these models now have intrinsic dynamical properties, which means that the internal state of the model evolves at its own rate as the various elements of information available to the model are integrated over time. Much effort has also been devoted

to developing models that are physiologically plausible and reflect how the brain interprets speech sounds processed by the auditory system. Yet another current trend in computational modelling is concerned with the mechanisms involved in the perception of connected speech, as opposed to isolated segments or words. The three focus papers (Grossberg, Gaskell, Greenberg) and three commentaries (Gow, Shamma, Tuller) included in this section examine how different aspects of speech perception can be accounted for by dynamical computational models.

2.4 Pathology

This section explores the potential connections between temporal integration in speech perception and developmental language disorders, particularly dyslexia and Specific Language Impairment (SLI). It has been claimed that these disorders may stem from a general, non-linguistic deficit in central auditory processing. According to the temporal processing deficit theory, as proposed by Tallal and her collaborators in a well-known series of articles[2], dyslexic and SLI children have difficulties in processing short and/or rapidly-changing speech (as well as non-speech) sounds. These difficulties have been interpreted as indicating that acoustic information in speech is integrated over a longer temporal window in dyslexic/SLI children than in control listeners. This theory has been widely debated, and many issues have been raised which relate to the potential causes of developmental language disorders, but also, and more generally, to speech and language processing in children. The focus paper (Rosen) and two commentaries (Bedoin, Mody) contained in this section critically assess the potential contribution of auditory processing deficits to the genesis of dyslexia and SLI.

2.5 Psychoacoustics

As the perceptual relevance of fine phonetic cues to phonological constrasts is now scrutinized in an increasingly large number of studies, phoneticians and speech scientists need a clear characterization of how speech sounds are represented in the auditory system. To assess the role of distributed phonetic properties in speech perception, we need to understand how sensitive the auditory system is to subtle spectral and durational variations in speech sounds, and how large

the window of integration can be in the time domain. We also need to better understand the auditory processes that may allow phonetic information to be simultaneously integrated over both short-time and long-time windows. More generally, the mechanisms involved in perceptual organization both for speech and non-speech sounds, such as auditory-stream segregation and fusion, need to be incorporated into current models of speech perception. These issues are addressed in the two focus papers (de Cheveigné, Moore) and two commentaries (Cooke, Macar) included in this section. The papers examine how auditory temporal integration is performed and how an internal representation of the input sound is generated. The role of temporal patterns in the auditory system above the cochlear level is also discussed.

2.6   Development

As shown by Peter Jusczyk, temporally-distributed phonetic properties are crucial to the development of speech perception and production. The main purpose of the two focus papers (Christophe, Gout, Peperkamp & Morgan; Port) and two commentaries (Best, Kello) contained in this last section is to explore temporal integration in speech perception from that perspective. The section centers on the infants' capacity to perceive phonetic information distributed across syllables, especially prosodic information, and examines whether this information may help infants to recognize lexical units. Another major issue is to what extent the sensitivity to prosodic information in infants can be accounted for by universal mechanisms governing the perception and production of rhythmic patterns.

3   Some common themes

3.1   The units of speech perception

A widespread view is that speech is decomposed by the listener into a sequence of basic perceptual units from which larger as well as smaller linguistically-relevant units are then retrieved[3]. Despite longstanding debate, there is no clear consensus as to which unit can be regarded as primary for the listener. Goldinger & Azuma point out that the list of proposed candidates (which include

features, phonemes and syllables) may have in fact grown over the years. Several contributions in this volume suggest that the problem may have been ill-posed, and that listeners are simultaneously sensitive to units of different sizes in the speech signal. In the Adaptive Resonance Theory (ART) proposed by Grossberg for example, smaller and larger units become simultaneously active, with a natural bias for larger units to prevail over smaller ones. Experimental support for this view is provided by Goldinger & Azuma, who demonstrate that by manipulating bottom-up and top-down sources of information appropriately, units of different sizes can be brought to the listener's consciousness is predictable ways. From a different theoretical perspective, that of Firthian Prosodic Analysis, Coleman, Hawkins and Local underline that the phonological contrasts relevant to the listener in speech perception may extend over different structural domains, as opposed to being confined to segmental contrasts, and, in addition, that both long-domain and short-domain phonetic properties may be associated with a phonological contrast.

3.2   Role of phonetic detail

Much attention has been devoted recently to the potential role of phonetic detail in the perception and understanding of speech. On the one hand, because speech perception is resistant to noise and to the huge intra- and inter-speaker variability of the acoustic signal, it has often been hypothesized that lexical access involves mapping speech onto a set of context-independent abstract features. On the other hand, recent research suggests that listeners are sensitive to phonetic detail, and retain many if not all the encountered phonetic variants of a word in memory. A prominent advocate of the first view is Remez, who contends that perceptual organization is indifferent to the details of sensory quality. His perceptual studies using sinewave speech indicate that listeners may focus their attention on patterns of variation in the gross spectral envelope, as opposed to short-term changes in fine-grained acoustic properties. Other contributors however, emphasize that phonetic detail may provide the listeners with crucial information in speech perception. Gaskell, for example, examines how assimilation in word-final coronals is perceived by listeners, and shows that the listeners' patterns of response can be accounted for by a probabilistic connectionist model exposed

8

to varying degrees of assimilation that reflect those encountered in real speech. Kello's concept of "temporal signature" refers to the fact that temporal patterns that are specific to a particular talker may serve to distinguish the speech signal emitted by that talker from other concurrent sources of sensory information. Kello provides interesting indications on how the listener's brain may use temporal signatures to perform source separation. Coleman, Goldinger & Azuma, Hawkins, and Local also argue that phonetic detail is perceptually relevant (see also Johnson (1997) and Bybee (2001) for similar viewpoints). Nygaard points out that both abstract time-varying patterns and fine phonetic properties may in fact be essential to spoken communication. Most exemplar-based models include a variety of mechanisms that allow more generic patterns to emerge as exemplars accumulate in memory. Likewise, an abstractionist model such as FUL, developed by Lahiri and Reetz (2002), includes a speech-processing front-end that performs feature detection using detailed acoustic criteria. These models therefore range on a continuum between a purely abstractionist approach and a "full-listing" exemplar approach to speech perception.

3.3   Form and function of phonological representations

To what extent do phonological representations come into play in speech perception and understanding? In the present volume, different answers are offered which partly depend on how the relationships between the phonetic and phonological levels of analysis are conceived. Local argues for a strict demarcation between these two levels, considered as being ontologically- and type-distinct. This may suggest that, as the speech signal is processed by the listener, an abstract and symbolic phonological representation is built up with respect to which phonetic exponents are interpreted. Coleman also emphasizes the abstract character of phonological contrasts such as [voice], which are said to be associated in an arbitrary fashion with large sets of distributed phonetic properties. (Thus, the darkness of the initial /l/ in *led* and the absence of aspiration in the final /d/, which both point to the stop being [+voice], are said to have no intrinsic connection with each other; for an alternative, perceptually-motivated view, see Hawkins and Nguyen (2004).) In the exemplar-based approach, in contrast, the sound-to-meaning mapping is performed through

myriad, continuous memory traces that bear a relation of natural analogy with the input auditory patterns. Consistent with this approach are proposals by Moore, Shamma and Greenberg that speech sounds are recognized by being compared with internal spectro-temporal excitation patterns (STEPs, in Moore's terminology) which are stored in long-term memory, based on previous experience. In that view, symbolic phonological representations may not be involved in some on-line processing of speech and may be sometimes computed after the meaning has been understood by the listener, as in Hawkins' Polysp model, for example.

3.4 Status of the lexicon

The status of lexical contrastivity in the perception of speech is another recurrent theme. Local considers that the central place typically attributed to the lexicon in theories of speech understanding has led to overemphasis on short-domain phonetic properties related to phonemic contrasts. Shifting the focus away from what he refers to as the "tyranny of the lexicon" allows Local to expose fine phonetic properties associated with phonological contrasts which are spread over long temporal windows, and/or that perform functions other than lexical differentiation. Thus, the phonetic exponents of phonological contrasts (including fine phonetic detail) are shown to differ in function words as opposed to content words, a phenomenon attributed by Local to the fact that function words and content words form two different systems of contrastivity, with a restricted inventory and less variation in phonological structure for function words. Local likewise demonstrates that fine phonetic detail is related in systematic ways to the time course of conversational interaction, and in particular to patterns of turn-taking and co-operative exchange of information in a conversation. Further support for this view is provided by Docherty, Hawkins and Laver, who underline that fine phonetic properties simultaneously perform several perceptual functions, and provide the listener with information about the socio-indexical aspects of speech as well as lexico-morphological information.

3.5 Development and learning

Yet another focus of interest for many contributors in this volume concerns the way in which phonetic/phonological representations come to emerge in the course of speech development, and to what extent these representations are shaped by how speech is perceived and produced. This issue was extensively examined in Jusczyk's work and led to the development of the WRAPSA model of word recognition and phonemic structure acquisition (Jusczyk, 1993, 1997). In accord with predictions made by WRAPSA, Christophe and her colleagues show that infants are especially sensitive to the prosodic structure of speech, which helps them to find word boundaries in a sentence. More particularly, they found that phonological-phrase boundaries are interpreted as coinciding with word boundaries by infants. Christophe et al. speculate that the infants' responsiveness to prosodic boundaries may be related to their sensitivity to periodic patterns. As pointed out by Port, the capacity to perceive and produce simple periodic patterns probably arises very early in child speech, and it may indeed provide a universal framework for the acquisition of phonology. Interestingly, Best suggests that prosody may also make it easier for infants to perceive phonetic information distributed over segments or syllables, and thus to partly overcome the fact that temporal integration is performed over a very short time window in infants, owing to their limited processing resources.

3.6 Auditory representations

The structure of auditory representations is dealt with by several contributors and is relevant to most papers in this volume. Psychoacoustic data allow speech scientists to better determine how detailed the auditory information available to the listener can be, and to identify the spectro-temporal patterns extracted by the auditory system from the signal. Both of these dimensions obviously condition how speech is represented at higher stages of processing up to the understanding of meaning. One important point made by both Moore and Cooke is that temporal integration does not amount to simply accumulating auditory information over a certain period of time, but instead involves combining information from multiple independent "looks" or "glimpses", that

coincide with regions of relatively high signal-to-noise ratio. These multiple looks may make the speech signal more intelligible to the listener in noise, and are associated by Cooke with Stevens' concept of acoustic landmark (Stevens, 2002). Vectors of subsequent looks are used to build what Moore refers to as a STEP, as indicated above (§ 3.3), and the STEP is in turn mapped onto stored templates by means of a pattern-matching process involving time warping. Similar hypotheses are developed by Greenberg and by Shamma. Another important point, made by de Cheveigné, is that temporal patterns (on the time scale relevant for pitch or timbre) may be transmitted beyond the cochlea, contrary to what has been previously assumed, and may be processed at a more central level. This is consistent with Macar's suggestion that the same neural structures are employed for temporal processing regardless of the sensory modality, and are also involved in the production of temporal motor patterns.

3.7 Methodological issues

Current models of speech perception become increasingly complex as they attempt, on the one hand, to begin with the speech signal itself (as opposed to what is referred to as "mock speech", that is an abstract, phonemic or featural representation of the speech input, sometimes derived from orthographic transcriptions) and, on the other hand, to go beyond the lexical level and account for how the meaning of an utterance is understood by the listener in a given communicative context. This results in an increase in the model's number of degrees of freedom, and thus in greater predictive power. It is therefore crucial to demonstrate that these models are falsifiable and lead to non-trivial predictions. These predictions must then be assessed by making relevant comparisons between the model's and the listener's behavior. These issues are addressed in different ways by several contributors in this volume. Tuller explains how, as she and her colleagues developed their dynamical model of speech perception, unexpected predictions made by this model were discovered that were then empirically tested. She emphasizes that models further our understanding of speech perception to a greater extent when they are counter-intuitive, and inconsistent with part of the possible outcomes in a given experimental situation. Gow underlines the limits of computational

models based on a schematic representation of the speech input, which by their very nature rule out the possibility of fine phonetic detail having a direct influence in lexical access. Gow rightly states that the issue of whether phonetic detail is or is not perceptually relevant is an empirical one, that has to be addressed in an empirical way, using computational models that are trained on real speech. Goldinger & Azuma show how the investigator's expectations can lead to the listeners focussing their attention on phonemes vs. syllables in a sequence monitoring task. Goldinger & Azuma's striking results bring to light the power of social influence on the listener's behavior, and they raise the question of to what extent the way in which the speech signal is interpreted by the listener depends on the experimental design. In a different area, Rosen and Mody closely examine the methodological problems associated with studying speech perception in dyslexic and SLI children.

4   Tribute to Peter Jusczyk

This special issue of the Journal of Phonetics is dedicated to the memory of Peter Jusczyk, who intended to give an invited paper at the TIPS workshop but tragically died before it took place. A eulogy for Peter Jusczyk is presented by Robert Remez.

Acknowledgements

Notes

[1]The abstracts of the focus papers, commentaries and poster papers were put together in a book published by the University of Cambridge printing service (ISBN: 1680-8908). This book is available on-line at `www.lpl.univ-aix.fr/~tips`. Two focus papers presented at the conference are not included in this volume, since their substance has been or will be published elsewhere. These papers were by Michel Habib on clinical intervention in dyslexia and Carolyn Drake on perception and development of rhythm and timing in music, examined intra- and inter-culturally. Descriptions of these papers are available through the TIPS book of abstracts and, to varying degrees, they are discussed in the commentaries. In particular, Maria Mody's extended critique of Habib's and Rosen's papers summarises Habib's main points, while Françoise Macar offers comments on Drake's paper as well as de Cheveigné's and Moore's. Finally, the volume does not contain four commentaries presented at TIPS, by Bailey (keynote session), Démonet (session on phonetic and phonological issues), Darwin (session on psychoacoustics) and Lindblom (session on development).

[2]See Rosen's and Mody's papers for references.

[3]Thus, syllables are seen in some theories as the primary units of perception, from which both words and phonemic units are identified.

References

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

Hawkins, S. and Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics*, 32. in press.

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

Jusczyk, P. (1993). From general to language-specific capacities: the WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21:3–28.

Jusczyk, P. (1997). *The discovery of spoken language.* MIT Press, Cambridge, Mass.

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.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer, Dordrecht.

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.