Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 46 (2005) 53–72
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception

Natasha Warner *, Roel Smits, James M. McQueen, Anne Cutler

Max Planck Institute for Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands

Received 18 December 2003; received in revised form 26 January 2005; accepted 27 January 2005

Abstract

We report detailed analyses of a very large database on timing of speech perception collected by Smits et al. (Smits, R., Warner, N., McQueen, J.M., Cutler, A., 2003. Unfolding of phonetic information over time: A database of Dutch diphone perception. J. Acoust. Soc. Am. 113, 563–574). Eighteen listeners heard all possible diphones of Dutch, gated in portions of varying size and presented without background noise. The present report analyzes listeners' responses across gates in terms of phonological features (voicing, place, and manner for consonants; height, backness, and length for vowels). The resulting patterns for feature perception differ from patterns reported when speech is presented in noise. The data are also analyzed for effects of stress and of phonological context (neighboring vowel vs. consonant); effects of these factors are observed to be surprisingly limited. Finally, statistical effects, such as overall phoneme frequency and transitional probabilities, along with response biases, are examined; these too exercise only limited effects on response patterns. The results suggest highly accurate speech perception on the basis of acoustic information alone.

Keywords: Speech perception; Diphone; Timing; Dutch; Feature

1. Introduction

Listeners' recognition of speech requires decisions which are phonemic in nature: for example, that a speaker said *bit* and not *sit*, *but* or *bill*. The identification of phonemic information to motivate such decisions, however, is affected by a multiplicity of factors beyond the acoustic cues which—invariantly or otherwise—directly signal

* Corresponding author. Present address: Department of Linguistics, University of Arizona, P.O. Box 210028, Tucson, AZ 85721-0028 USA. Tel.: +1 520 626 5591; fax: +1 520 626 9014.

E-mail addresses: nwarner@u.arizona.edu (N. Warner), heersmits@hotmail.com (R. Smits), james.mcqueen@mpi.nl (J.M. McQueen), anne.cutler@mpi.nl (A. Cutler).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.01.003

phonemic identity. Thus identification responses are affected by the surrounding phonetic context in which a phoneme occurs, by the phoneme's position in a word or utterance and consequent differences in prosodic realization, as well as by listener expectations based on past experience, as when phoneme frequency effects or transitional probabilities play a role. Decades of speech perception research have been devoted to exploration of these factors (see Nygaard and Pisoni, 1995, for a review).

We here report analyses of these effects in a very large database of perceptual identifications. In speech research, very extensive databases have enabled important advances in our knowledge. Thus Miller and Nicely's (1955) database of perception of consonants in noise, Peterson and Barney's (1952) database on vowels and the subsequent work of Hillenbrand et al. (1995), as also the segment and syllable duration data of Crystal and House (1982, 1988a,b) have all proved treasure-houses for scholars working on a range of speech-related topics. Such extensive databases allow for comparison of many factors with experimental methods held constant, so that the information provided is directly comparable across segment types, stress positions, etc. The database which we describe here concerns perception of segments in Dutch in all possible immediately adjacent contexts. Collected via a gating task, the database gives a temporal view of how Dutch listeners perceive the sounds of every diphone (two-phoneme sequence) in the language, as acoustic information becomes available with each gate.[1]

The choice of diphones as the test set was motivated jointly by considerations of validity and feasibility. For validity, phonemic identification must be assessed in context. Clearly, the goal which listeners aim for in speech recognition is not apprehension of a sequential representation of phonemic units. Listeners want to know what the speaker wished to communicate, i.e. they are interested in meaning, and hence in recognizing the words which comprise an utterance. Phonemes are crucially relevant not because they are an end in themselves, but because they constitute minimal differences between words such as *bit* and *sit* or *but* or *bill*. We therefore wished to examine the uptake of phonemic information in all possible contexts. The larger the context, the better; but even triphone sequences would have presented us with a set of tens of thousands of stimuli, so on grounds of feasibility of data collection we chose diphone sequences. (Even then, there were over a thousand such possible sequences, and by varying stress and presenting the diphones in fragments of varying size, we ended up requiring our listeners to respond to over thirteen thousand stimuli, which took on average 27.9 test hours per listener.) Diphones thus offered the minimal contextual environment for a feasible study of natural perception of phonemic information in speech.

The database itself is publicly available: http://www.mpi.nl/world/dcspdiphones. Smits et al. (2003) describe in detail the methods used to collect the database. That methodological report contained however only the most summary statistics concerning the perceptual findings, namely percent correct judgments per gate for segments individually, and averaged across consonants, across vowels, and across all segments.

The data reported by Smits et al. (2003) nevertheless showed clearly how listeners progress in their perception of sounds, both for the first and the second sounds of a diphone. The most important patterns which Smits et al. observed for consonants were: (1) Stops were not recognized well until listeners could hear their bursts. (2) Voiced obstruents (both stops and fricatives) tended to be misperceived as the voiceless equivalent, but the confusion did not go in the opposite direction. (3) Fricatives could be recognized very well from the first third of the fricative, but not from the preceding vowel, so that improvement in perception of fricatives (both voiced and voiceless) was quite sudden at the first gate that included frication noise. (4) Useful information for perception of nasals was available both in the final portion of the preceding sound and, even more so, in the first

[1] Responses in the gating task, of course, represent listeners' conscious decisions about what sounds they have heard, rather than their online recognition of sounds as a part of spoken word recognition. See Norris et al. (2000) for extensive discussion of this distinction.