SPEECH PERCEPTION

Spoken-Word Recognition

Anne Cutler

Max Planck Institute for Psycholinguistics,
Nijmegen, the Netherlands

The psycholinguistic branch of speech perception research is principally concerned with the study of spoken-word recognition. As a separate area of research, spoken-word recognition boasts a scant quarter-century of history. Nonetheless, the field has grown and changed in many ways dining that time. This brief essay makes no attempt to survey the field, or to isolate its principal current controversies; instead, the emphasis is on extrapolation to the future. Three current trends in the field are described, with the expectation that these are continuing developments which are likely to make further progress in the coming few years.

Trend 1. Greater computational explicitness in modelling

Psycholinguistic studies of the recognition of spoken words began in the 1970s; this decade was notable for the development of tasks specific to auditory word recognition, such as phoneme-monitoring (Foss, 1970), gating (Ellis et al., 1971), mispronunciation detection (Cole, 1978), cross-modal priming (Swinney, 1979) and phoneme-categorisation in lexical context (Ganong, 1980). These tasks made possible a substantial increase in the proportion of auditorily-based studies in psycholinguistic research on language comprehension. At the same time, researchers realised that the large body of results accumulating for visual word recognition did not necessarily illuminate the spoken-word recognition case, because of the temporal nature of speech signals.

Probably the most explicit statement of the need for separate modelling of spoken-word recognition came with the proposal of the Cohort model by Marslen-Wilson and Welsh (1978). This model embodied testable claims about the recognition of spoken words, and prompted a good deal of research. Likewise, the Logogen model (Morton, 1970), although not specific to auditory recognition, prompted much research because it made testable predictions. However, these predictions were rather general in nature (words are recognised left-to-right - a claim of the Cohort model; effects of frequency of occurrence and of semantic context are additive - a prediction of the Logogen model); the generality reflected the fact that it was not possible to derive predictions of an explicit, and, in particular, quantifiable, nature from either model. Neither the Cohort model nor the Logogen model was ever computationally implemented. This was in fact equally true of other models of the 1970s (attempts were made to implement the LAFS model [Klatt, 1979] but without success; see Klatt, 1989). Modelling of spoken-word recognition changed greatly in the late 1980s. The connectionist model TRACE (McClelland and Elman, 1986) was the first widely influential computationally explicit model in this area. The Fuzzy Logical Model of Perception, or FLMP (Massaro, 1987), is similarly explicit (note that this model is not supposed to apply solely to speech perception, although most of the work within the FLMP framework has in fact been in this area). The Neighborhood Activation Model, NAM (Luce et al., 1990), represents the interactions between words of greater and lesser phonological similarity, and greater and lesser frequency, in the vocabulary, i.e. it models word recognition as a function of the makeup of a language's stock of words.

The NAM has been implemented for a subset of the English vocabulary - words of CVC structure - and experimental results accord with the model's predictions (Goldinger et al., 1989; Luce et al., 1990). TRACE, too, has been implemented with a small vocabulary of just a few hundred words, and indeed only a subset of the phonemes of English, but has been able to generate predictions which have accorded with the outcome of experimental tests (McClelland and Elman, 1986; Elman and McClelland. 1988).

The obvious next step was computationally explicit models which could be implemented with a realistically-sized vocabulary, and of these there is so far just one available: Shortlist (Norris, 1994). Shortlist, a connectionist model, can run on a vocabulary of some 26,000 words. Such a large vocabulary would, with currently available computing resources, make any of the above models run very slowly indeed. Shortlist can handle the large vocabulary because it is a hybrid model, in which an initial stage selects, on the basis of the bottom-up input information, potential word candidates from among the entire available vocabulary, but the final word recognition process operates (on the basis of inter-word competition) among only the subset (the "shortlist") pre-selected in that initial stage. This model, too, has generated empirical predictions which have been tested and supported in laboratory studies (McQueen et al., 1994; Norris et al., 1995; Vroomen and de Gelder, 1995).

No psycholinguistic model of spoken-word recognition as yet accepts real speech input, although the proponents of some models have given serious consideration to how their work could be extended in this direction (Elman and McClelland, 1986; Elman and Zipser, 1988; Norris, 1990). Future developments are likely to bring progress in this respect. What is certain, however, is that modelling of spoken-word recognition will remain computationally explicit. It is now quite common for research in this area to involve an interplay between experiment and explicit predictions of an implemented model, combined with actual simulations; the empirical papers cited in the above two paragraphs are all examples of this. This could indeed become the rule.

Trend 2. A realistic view of the task

This trend is in fact not fully separate from the increasing explicitness of modelling. The availability of lexical databases on computer has made it possible to incorporate the characteristics of a real vocabulary into modelling efforts. The vocabulary of the average language user comprises many tens of thousands of words, so that taking the contents of the vocabulary into account, especially in a computationally implemented model which can simulate recognition with a large vocabulary, greatly improves modelling realism. Such lexical databases have now been available for many years and have been exploited in theoretical studies of spoken-word recognition since the mid-1980s (see e.g. Luce, 1986; Marcus and Frauenfelder, 1986). What is now becoming available as the next step is computerised corpora of real speech. The London-Lund Corpus (Svartvik and Quirk, 1980) was available only in orthographic transcription, although despite that limitation it did offer some useful data for speech perception research (e.g. Brown, 1984; Cutler and Carter, 1987). Rut much more useful will be speech corpora for which the original speech is accessible; one such corpus is MARSEC (Roach et al., 1993), which is a machine-readable version of the Spoken English Corpus of about 55,000 words of British English speech.

A further sign of increasingly realistic word-recognition research is the renewed attention to cross-linguistic comparisons in this field; languages do not all present exactly the same set of problems to be solved in a recognition model. Although the early days of psycholinguistics saw a good deal of cross-linguistic research, the research climate of the 1970s (when spoken-word recognition came into its own) did not favour comparative approaches. More recent studies (e.g. Cutler et al., 1992; Sebastian-Galles et al., 1992) consider language structure - language-specific phonological structure, in the case of speech perception - as part of the subject matter of the field. An assistance to such approaches has been the availability of computerised lexical resources not only for English but for other languages (e.g. CELEX, which combines data for Dutch, German and English: Baayen et al., 1993). We should expect further increase in cross-linguistic studies which use characteristics of a particular language or vocabulary to illuminate theoretical questions pertaining to a universal model of spoken-word recognition.

Corpora such as CELEX and MARSEC are easily accessible because they arc available on CDROM; one can expect further exploitation of such resources as more of them become available in this format.

Trend 3. Towards more "direct" methodologies

Traditional techniques of psycholinguistic investigation involve laboratory presentation of controlled input and measurement of some response on the part of the listener. In the field of spoken-word recognition a high priority has been attached to achieving insight into recognition "on line", i.e. into charting the time-course of word processing. No direct window into mental processing being as yet available, so-called on-line techniques have involved the measurement of response time to perforin some task - either word recognition itself (as in the lexical decision task, for example), or some task which putatively involves a component of word recognition (such as detection of a target phoneme) or putatively depends upon recognition (such as word repetition). All of these tasks are necessarily indirect, and involve inferential steps from the response time to the proposed implications for our knowledge of the recognition process, including partialling out the components of the response contributed by attentional and decision aspects of the task.

A revolution which is taking place in psycholinguistics, and is sure to make itself felt in spoken-word recognition research, is the advent of techniques in which some measurement is made of direct cortical response to language input. Such measures are assumed to avoid the attention and decision components of response-time tasks, and to offer a more direct view of linguistic processing. In fact the techniques are very much still in their infancy and it is perhaps illusory to think that they are as yet in a position to provide a direct picture; in particular, the reasoning connecting the characteristics of the input to the measured response certainly still involves many inferences. However, these methods are less indirect than traditional techniques in the sense that no part of the response appears to be under the listener's voluntary control; and we can certainly expect much progress in the sophistication of cortical-response measurement in the coming years.

For spoken-word recognition, the two methods which initially offer themselves are Positron Emission Tomography (PET) and Event-Related Potential measurement (ERPs). These two techniques offer rather different views of processing, in that PET has relatively good spatial resolution but poor temporal resolution, while ERPs have good temporal resolution but less than optimal spatial resolution.

PET involves an injection of radioactively labeled fluid, and scanning of the brain (the most advanced machines can take dozens of cross-sectional images) to assess (changes in) regional cerebral blood flow. It is, as can be seen, an invasive method, and it is therefore not always easy to justify its use as a basic-research tool with normal listeners; much research on language processing using PET is carried out on patients who are undergoing a scan for medically imperative reasons. The accuracy of the results from a PET scan can be improved by combining it with a series of high-resolution magnetic resonance images (MRI) for each subject; this enables variations in blood flow to be assigned to cortical structures individually defined for each subject. (Functional MRI as a measurement technique itself has at present limited applicability to the study of speech processing simply because of the high noise level of MRI devices. Future developments may overcome this limitation; if so, functional MRI, less invasive than PET, may offer challenging basic-research possibilities.) Demonet et al. (1994), and Mazoyer et al. (1993; this paper also involves cross-linguistic comparisons) exemplify use of PET with auditory linguistic input.

FRP measurement is, in contrast, a non-invasive method, and is suited for use with all subjects. (In fact, state-of-the-art high-density ERP measurement has been used with very young infants; Dehaene-Lambertz and Dehaene, 1994). Current techniques involve a large number of electrodes - 32- to 128-electrode systems are standardly in use - placed on the outside of the scalp, usually mounted in a cap, with additional electrodes to provide reference levels. Responses may be averaged across intervals (e.g. of a tenth or a quarter of a second). Holcomb and Neville (1990) and Aaltonen et al. (1994) are studies representing respectively more psycholinguistic and more phonetic uses of the ERP technique.

These techniques will not replace behavioural measurement in psycholinguistic studies of speech perception, but they will increasingly augment them and provide converging evidence with more traditional tasks; like the increasing explicitness and the increasing realism of psycholinguistic models, this is a current trend which seems likely to progress further in the immediate future.

References

Aaltonen, O., Eerola, O., Heikki Lang, A., Uusipaikka, E., and Tuomainen, J. (1994). Automatic discrimination of phonetically relevant and irrelevant vowel parameters as reflected by mismatch negativity. *Journal of the Acoustical Society of America,* 96, 1489-1493.

Baayen. R.H., Piepenbrock, R., and van Rijn, H. (1993). *The CELEX lexical database (CDROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instrumentation and Computers,* 16, 502-532.

Cole, R.A. (1973). Listening for mispronunciations: a measure of what we hear during speech. *Perception and Psychophysics,* 11, 153-156.

Cutler, A. and Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer, Speech and Language,* 2, 133-142.

Cutler, A., Mehler, J., Norris, D., and Segui.J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology,* 24, 381-410.

Dehaene-Lambertz, G. and Dehaene, S. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature,* 370, 292-294.

Demonet, J-F., Price, C, Wise, R., and Frackowiak, R.S.J. (1994). A PET study of cognitive strategies in normal subjects during language tasks: Influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain,* 117, 671-682.

Ellis, I... Derbyshire, A.J., and Joseph, M.E. (1971). Perception of electronically gated speech. *Language and Speech,* 14, 229-240.

Elman.J.L. and McClelland,J.L. (1986). Exploiting lawful variability in the speech wave. In J.S. Perkell and D.H.. Klatt (Eds.): *Invariance and variability of speech processes,* 360-380. Hillsdale, NJ: Erlbaum.

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

Elman.J.L. and Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America,* 83, 1615-1626.

Foss, D.J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision *times. Journal of Verbal Learning and Verbal Behavior,* 8, 457-462.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance,* 6, 1 10-125.

Goldinger. S.D., Luce. P.A., and Pisoni, D.B. (1989). Priming lexical neighbours of spoken words: Effects of competition and inhibition. *Journal of Memory and Language,* 28, 501-518.

Holcomb, P.J. and Neville, H.J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes,* 5, 281-312.

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

Klatt, D.H. (1989). Review of selected models of speech perception. In W.D. Marslen-Wilson (Ed.): *Lexical Representation and Process,* 169-226. Cambridge, MA: MIT.

Luce, P.A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics,* 39, 155-158.

Luce, P.A.. Pisoni, D.B., and Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T.M. Altmann (Ed.): *Cognitive Models of Speech Processing,* 122-147. Cambridge, MA: MIT Press.

Marcus, S.M. and Frauenfelder, U.H.. (1985). Word recognition - uniqueness or deviation? A theoretical note. *Language and Cognitive Processes,* 1, 163-169.

Marslen-Wilson, W.D. and Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology,* 10, 29-63.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Erlbaum.

Mazoyer, B.M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier. O., Salamon, G., Dehaene, S., Cohen, L., and Mehler.J. (1993). The cortical representation of speech. *Journal of Cognitive*

Neuroscience, 4, 467-479.

McClelland, J.L. and Elrnan.J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology,* 18. 1-86.

McQueen, J.M., Norris, D.G., and Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 20, 621-638.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G.T.M. Altmann (Ed.): *Cognitive Models of Speech Processing,* 87-104. Cambridge, MA: MIT.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition,* 52, 189-234.

Norris, D.G., McQueen, J.M., and Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 21, in press.

Roach, P., Knowles, G., Varadi, T., and Arnfield, S. (1993). MARSEC: A machine-readable Spoken English Corpus. *Journal of the International Phonetic Association,* 23, 47-53.

Sebastian-Galles, N., Dupoux, E., Segui.J., and Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language,* 31, 18-32.

Svartvik, J. and Quirk, R. (1980). *A Corpus of English Conversation.* Lund: Gleerup.

Swinney, D.A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior,* 18, 645-659.

Vroomen.J. and de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance,* 21, in press.

Hot Topics in the Field of Speech Perception?

Sieb Nooteboom
Research Institute for Language and Speech,
Utrecht University, the Netherlands

I have been asked by the editors of this book to give my "views on the current and future hot topics in the field of speech perception". Not an easy thing to do. The field of speech perception today is a broad area, covering widely divergent research questions, investigated by researchers with very different backgrounds and very different goals in mind. Because of this variety of goals and interests, pursued by a limited number of researchers, research efforts are thinly spread, and, at this moment in time, it seems to me an exaggeration to call any one topic in this area "hot". However, in the past there have been several "hot topics" in the domain of speech perception. In the fifties and early sixties, there were a number of researchers who firmly believed that "speech is special", and found evidence for this is in a great number of experiments showing or attempting to show that speech sounds, and particularly consonants, are perceived categorically in a way that non-speech sounds are not (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Although there was no lack of publications refuting the specialty of speech, the issue was revived in a somewhat different way some twenty-five years ago. Between 1970 and 1980, many more than 100 articles were published on the single hot topic of hypothesised, possibly innate, feature detectors in the human brain specialised for distinctive features in speech (for relevant literature see Cooper, 1979). Researchers attempted to find evidence for such feature detectors mostly by way of the method of "selective adaptation". It was believed that one could fatigue the feature detectors by subjecting them to overstimulation. The fatigue was hoped to show up in a shift in a perceptual criterion. In the end empirical evidence became more and more confusing, and the quest for linguistic feature detectors turned cold. Since then, very few people still