Vernieuwingsimpuls / Innovational Research

Grant application form 2003 VIDI scheme

Please refer to Explanatory Notes when completing this form

Registration form (basic details)

1a. Details of applicant

-Name, title(s): R.J.J.H. van Son, Dr. (Rob)
-Male/female: Male
-Address for correspondence:

1b. Title of research proposal

Integration of information in spoken communication

1c. Summary of research proposal

(max. 300 words, plus max. 5 KEYWORDS)

Our understanding of the comprehension of spoken language is lacking on quantitative knowledge on how the different aspects of language are integrated. Both the time-course with which information becomes available and the way the diverse sources of information are combined are relatively unknown. Speech recognition in the classical sense of "structured word-recognition" is an extremely complicated process. It is necessary to start tackling the general problem of the extraction and integration of information in speech comprehension with a simpler sub-task. A much simpler problem, which covers the whole spectrum of language communication, is the prediction of turn-switches in conversation. Turn-switches in various forms are the basic control mechanism of conversations. For the hearer, the task is deceptively simple: determine when to start talking. This makes turn-switching a good model for the extraction and integration of linguistic information as all sources of relevant information are synchronized with the turn-switching points (Turn-Relevant-Places or TRP's). From an experimental point of view, the interference from the task itself, whether or not to start speaking, is minimal, as the number of choices is extremely limited. Therefore, the research can concentrate on the integrating process itself. The proposed project concerns the quantitative modeling of TRP identification in conversation as an integration process of temporally unfolding information at different levels in speech, from conversation-acts and semantics to prosody, phonetics, and visual cues. Reaction Time (RT) measurements from TRP monitoring in manipulated (partial) conversations will be used to determine exactly when the relevant information at different levels of speech becomes available and how it is integrated to predict the position of a TRP. We will especially look at generalizations of the MERGE model extended with a Random-Walk decision model. We will include both the standard flat Bayesian decision rule and more structured Hierarchical models of integration.

Key words: speech comprehension, information integration, conversation, talk-in-interaction, turn-switching

1d. NWO Council area

GW

1e. Host institution (if applicable)

Leerstoelgroep Fonetische Wetenschappen/Onderzoeksinstituut ACLC
Afdeling taal- en letterkunde
Faculteit der Geesteswetenschappen, Universiteit van Amsterdam

Vernieuwingsimpuls / Innovational Research

Grant application form 2003 VIDI scheme

Please refer to Explanatory Notes when completing this form

Research proposal

2. Description of the proposed research

(max. 4000 words on max. 8 pages, excl. literature references. A description of sub-projects is required. Include details of:)

2a. Research topic

Understanding language comprehension is crucial for understanding human cognition and social life in general and human communication in particular. Within this wider field, *spoken* language is further characterized by a strong time-pressure on comprehension. An utterance has to be processed before the next one supersedes it. The classical view of spoken language is based on the structure of written texts [24]. In this view, speech is considered to be a linear "string" of regular units that encodes a corresponding sequence of meanings. In the modern view, speech is understood as a lattice of interconnected, hierarchically structured layers consisting of “streams” of, e.g., phrases, words, and phonemes. Moreover, utterances carry multiple, concurrent messages beyond the “literal” (textual) meaning encoded in the words. These messages include group membership, emotional state, and conversational control signals. For each of these concurrent messages, listeners integrate many of the streams to extract the relevant content from the speech. The same sounds and visual cues that are used to recognize the words also give an accurate picture of the emotional state and socio-linguistic background of the speaker. Furthermore, the listener will be acutely aware when and how she is expected to react, be it with back-channel utterances, nods, or to take her turn in the conversation.

For each of the theoretically recognized aspects of a conversation, there are successful models on how it is structured (cf., [2, 7, 26, 27, 34]). However, our understanding of how the various parallel streams of language are integrated in comprehension is lagging behind. Three questions are central to understanding the processing of language in recognition. *What* information is available, *when* does it become available, and *how* is it used? We have a reasonable picture of the information available in speech. However, our understanding of the time course of processing is, in general, rather superficial, although the study of (cross-modal) priming has clarified many processes (e.g., [11]). When it comes to the question of how the information is used to integrate all the different aspects of an utterance into a single “meaning”, we have only a very crude understanding on how to synchronize and integrate the individual layers of information.

It is clear that the above questions in their general form are currently intractable. The answers tend to depend on the (sub-)task facing the listener. The mechanisms used to integrate diverse sources of information from *speech* can be illustrated on two examples, the short-range McGurk effect and the long-range Ganong effect.

In the McGurk effect [20, 19], seeing a movie of a face that utters a stop-consonant with a velar/uvular place of articulation (e.g., /g/) can alter the perception of a stop-consonant plosive sound with a labial place of articulation (e.g., /b/) that is heard simultaneously, to a /d/. This outcome is determined by a weighting of visual and auditory cues. The integration of visual and acoustic cues is very fast, before the phoneme is classified.

In the Ganong effect, an ambiguous instance of a plosive can be recognized as either a voiced or an unvoiced version of this plosive, depending on the context in which it is presented [4]. So in the sentence 'How to milk a coat', an ambiguous k/g sound will tend to be perceived as the /g/ of 'goat' whereas in another context, e.g., 'When it rains I put on my coat', the same ambiguous sound will tend to be perceived as a /k/. In the Ganong effect, we see that a long range context, in the order of seconds, can affect the interpretation of an ambiguity that is based on a 10 ms difference in voice-onset time. Simplified, the ambiguous phoneme activates two words in the lexicon, 'goat' and 'coat'. Only one fits the meaning of the sentence and this one will 'bias' the lexical selection and the corresponding consonant is 'heard'.

Some of the most successful experimental paradigms in elucidating the time-course of human language processing are based on Reaction-Time measurements (RT), e.g., reading time, phoneme monitoring, word/non-word classification, and priming studies in general. Most RT-paradigms are based on lexical processes. When applicable, RT studies are the most sensitive probes into the processing demands on subjects and the use of contextual information (e.g., [25, 11, 26]). Therefore, RT experiments are the best paradigm to study whether (and when) concurrent information is actually used and combined in solving a language task. In addition, as this project targets the integration of *all* information, including non-verbal information, the experimental paradigm must depend as little as possible on lexical processes and also must synchronize different levels of language description.

One language task that is independent of lexical access and at the same time integrates many levels in speech is turn-switching in conversations. Turn switching is a natural and frequent part of any conversation. It happens around well defined time-points in speech where all aspects of speech indicate (predict) a turning point. It is fairly easy to elicit turn-related judgments from naive listeners, be it with RT or shadowing experiments, or with off-line annotation. Manipulated speech, facial-video's, or even texts, can be used to add or remove specific cues and look at their relative importance. It is well known that turn-switching is signaled at all levels of an utterance. The exact TRP is indicated by visual cues (gaze direction, e.g., [28]) and by specific intonational movements and phonetic changes in duration and loudness and the presence of silent pauses. The precise start of the new turn with respect to the TRP is an important message in itself [30]. In short, turn-switching is a tractable and accessible feature of speech that integrates many important aspects of language use.

An example from Schegloff illustrates this [28]. On page 410, a short conversation fragment is printed. It contains the following two turns (Before, *another* speaker has asked for the 'butter'):

Nancy: C'n I have some t[oo
Michael: [mm-hm-hm ...

Nancy's turn ends right after *too* where the question that started with *Can I* is completed. An earlier completion point (TRP) could be after *some* as the word *too* is redundant. However, un-transcribed aspects of the word *some* in this example, e.g., gaze direction, intonation, local speaking rate, and phoneme reduction, might have indicated that the TRP had not yet been reached. The next speaker, Michael, signals the redundant nature of *too* by starting his (ambiguous) turn right after its onset. His interruption point is indicated by the [-brackets, which align the turns (timing is very imprecise in this transcription). This early start might indicate that Michael predicted the word *too*, and did not wait for its completion. Another possibility is that Michael indeed reacted to a TRP right after *some* and ignored *too*. An analysis of video-recordings, intonation, and phonetic structure might elucidate whether there was a TRP projected after *some*, which was followed by extra material [1].

The presence and projection of TRP's can be studied in (naive) human subjects in several ways, e.g., in a straightforward *press when you would take turn* RT task or by some more sophisticated laryngograph measurement in a 'shadow' task. Subjects would hear Nancy's utterance, e.g., unedited, gated, monotonized, or in a "hummed" version. Differences between the subject's responses for all these manipulations would point out how the information on the upcoming TRP builds up in time in the separate "channels" of speech, e.g., words, intonation, and phonetic structure, and how they are combined.

From the perspective of the next speaker, three time-points are important with respect to TRP's. The first is the point at which the new speaker realizes that a TRP is imminent and she should start preparing a turn. In the above example this preparation point might be located after *have*. The second is the point where the next speaker expects the TRP to be realized. This could be located at the recognition points of *some* or *too* in the example. And third, the time at which the turn is actually taken by the next speaker. This point is indicated by the [-bracket in the example. Only the actual turn-taking is readily visible in the recorded conversation. Still, the timing of subject responses can be used to estimate the two earlier points.

Currently, the best quantitative model for the on-line integration of information in RT experiments is the Merge model [21] extended with a Random-Walk decision model [22]. The extended Merge model predicts reaction times as a stochastic summing in time (Random Walk model) of information from independent channels (Merge model). It can be generalized to include other information sources beyond phoneme recognition and lexical access and address TRP decisions instead of phoneme monitoring. The current version of the Merge model uses a flat Bayesian decision rule, which is a kind of default model (e.g., [19]). However, semantic information allows much less temporal precision than, e.g., phonetic information. So we can also propose a second, hierarchical model, where *long-range* effects, like semantics and syntax, predict a broad interval and *short-range* effects, like phonetic effects, are used to specify the exact point within this interval.

To summarize, we propose to use turn-switching as a tractable model system for the integration of linguistic information in speech recognition. Naive, native subjects are presented with partial and manipulated conversations and are asked to monitor for turn switches or back channel utterances using RT and shadowing paradigms (e.g., using laryngography or voice detection to time subject responses). We compare the delays with respect to the theoretical TRP's and the original turn-switches to study the underlying processes involved in *comprehension* and *decision making*. Many techniques are available to manipulate the available speech (and video) signal. For instance, PSOLA resynthesis can be used to manipulate prosody and intonation [3]. There are many techniques to make utterances (partially) unintelligible while preserving prosody and pause structure. Visual information can be presented in isolation, together with original or manipulated sound, or not at all. More elaborate designs, like eliciting real dialog responses from subjects or recording evoked potentials in subjects, are currently too ambitious.

PhD student sub-projects

The proposed PhD project requires a balance between challenge and predictability for a graduate student (AIO). The PhD project will address the question of how TRP's are projected by local, short-range signals like gaze-direction and acoustic parameters (both in production and perception). In this project it should become clear how listeners pin-point the precise position of an upcoming TRP from the local intonational movement, prosody, and acoustics of the utterance, and how speakers guide them in this. What little is known about this is based on intonation and prosody in general [1, 17, 36], pauses, and timing. Even less is known about the use of "tempo" with respect to intonation [1, 17, 8]. The work of Caspers [8], Auer [1], and Koiso et al. [17] will be good starting points for studying prosody in general. This can then be extended to include pronunciation variants and reduction.

2b. Approach
I will start this section with stressing the focus of this project:
Conversation participants communicate Turn-Relevant-Places. The proposed project wants to determine:

- *What* information about the upcoming TRP is available in the speech? (including other modalities)
- *When* is the relevant information available?
- *How* is this information used in discourse?

These questions are intricately linked and will be investigated using the MERGE model with a Random Walk decision model [21, 22]. Information use will be modeled using both hierarchical and flat (Bayesian) decision rules. In a hierarchical decision model, lower-level (short-range) information can only be used in the presence of higher level (long range) information, e.g., in attentional filtering. In a flat, Bayesian, model, all information is summed, weighted by their predictive value. A hierarchical decision model predicts sudden changes in the use of short range information when long range information is suppressed, whereas flat models predict very smooth changes and compensation between levels.
There are affinities between the proposed project and the Cognition program (e.g., *The neurophysiology of communication*) and the IMIX program (e.g., *Dialoogmanagement en redenering*). I will contact these programs before the project starts. I have already contacted scientists with expertise in the different "aspects"of language and conversation to supplement our own knowledge. To fit this study into a single VIDI project, the scope of the subject matter has been limited. The main limitation I adopt for a 5 year project to be successful is that only off-the-shelf models of language processing will be used. Whenever possible, existing annotated Dutch speech corpora will be used (e.g., the *Spoken Dutch Corpus*, CGN [23], and the Map Task, [8]. Any additional annotation will be constructed using established semi-automatic methods, from conversations-act theory and semantic structure down to phonemic segmentation (e.g., [32]; Van Son et al., 2001). The practical side of this project is assembled from the standard, off-the-shelf procedures used in the research of conversation and discourse analysis, semantics, syntax, prosody, visual speech correlates, and acoustic variability (e.g., [1, 17, 18, 27, 25, 28, 30, 33, 37, 36] and references therein). The focus of this project will be on the use of the information used to project and identify TRP's: *What* is available to conversation participants, *when* is it available, and *how* is it used to predict an upcoming TRP? This explicitly includes visual, durational, and spectral measurements around turn-switching moments (e.g., [28, 29]). The acoustic analysis is comparable to the measurements done to correlate acoustical parameters with break-indices in prosodic research [7, 12, 13, 15, 35, 38].
Audio and video recordings of Dutch conversations are scanned for TRP's according to the standard procedures [34, 9]. The surrounding utterances are isolated and analyzed on syntax, prosody, and phonetic and visual cues in the conventional way, e.g., gaze direction and face orientation (when available), syntactic completion and boundary tones. Speech around TRP's is segmented at the phonemic level as much as possible. Annotation, labeling and segmenting have to be partly outsourced as these contain highly specialized tasks.
This project combines three domains. *First*, it analyzes the structure of conversations. On the very practical side, I will use the annotation formalism and tools developed by Henry Thompson and his group on the large Map-Task database at the HCRC in Edinburgh [34, 9]. In the Edinburgh formalism, speaker turns are distributed at the so-called MOVE level. Speech act segmentation and coding can be done reasonably reliable by instructed naive (and native) coders [9]. The Edinburgh group makes their annotation methods readily available (personal communication from Amy Isard).
The second domain is the projection of TRP's by the speaker. We use turns as an observable behavior [34]. Determining the exact TRP point in time is based on theoretical models of the ideal speaker/listener (cf., [14]). A TRP is determined by the actual MOVE type, syntactic completion, local speaking "rate", voice quality, gaze direction, and boundary tone. This can easily be generalized with syntactic extensions and more general prosodic/temporal structures (e.g., [1, 17, 10]). There is a lot of experience with the analysis of conversational boundary tones in the Netherlands [8].
The third, and central, domain is the processing of all the information by the listener, both real and experimental participants to the conversation. The aim is to determine how and when subjects perceive an upcoming TRP and how and when they decide the TRP is there. The three time-points discussed earlier, i.e., the moment an imminent TRP is predicted, the time-point where it is projected to appear, and the time point it is actually realized, are measured both in the original conversation and in an experimental setting. The latter two points in time can be studied with straightforward RT and Shadowing tasks on original and manipulated recordings. For

the last, observable, turn-taking point, this would be *Push button when a turn is taken, Repeat the next speaker,* or *Speak 'eh' when the next speaker starts*. For the predicted turn-switching point this would be *Push button when you would take a turn* or *Say 'eh' when you would take the turn* (*eh* is the universal *take-the-floor* signal). The standard *Push button* paradigm has the advantage of being well understood. Shadowing has the disadvantage of interference from speech production. However, speaking a new turn instead of pushing buttons, is a more natural, and faster, automatic, behavior. Both methods will measure a combination of realized and predicted TRP. Neither will inform us about the preparation point where it is realized that a TRP is imminent. No tested methods are available to measure this point. However, we know that speakers tend to hold their breath some 200 ms prior to starting to speak by closing their glottis. This preparatory closure can be measured with a laryngograph in our experimental subjects [16]. So we will do a *Say 'eh' when you would take the turn* type task with laryngograph recordings using naive subjects. We expect to see a signal prior to the actual 'turn-taking' which indicates that the speaker prepares an utterance. This glottal signal will also show less interference from speech production than the actual utterance. There is ample experience with laryngograph recordings at the Chair of Phonetics and earlier on in the ESPRIT "SAM" Project (No 2589).
The time course of processing is modeled using RT-type analysis, e.g., the MERGE model and Random-Walk models [21, 22]. Integration of information is studied from different viewpoints like efficient decision theories (e.g., forecasting TRP's using Bayes' statistics, Neural Networks, CART trees, Minimal Message Length Coding), game theory (e.g., maximizing "gains" of the participants like getting the next turn), and minimizing repair costs. Non-symbolic aspects of conversation, e.g., gaze direction and speaking rate, will also be studied using ethological approaches [5, 31].
This study will be mainly build on existing material, annotated Dutch CGN and Map Task recordings ([23, 8]), and new (video-)recordings. As a start, a sub corpus and database of all observable TRP's in the annotations will be constructed, i.e., take-turns, hold-the-floor, grab-the-floor, back-channels. Initially, about 1000 TRP's will be selected for human inspection, to be extended later in the project. These will form a core corpus for detailed segmentation and acoustic analysis. From these, a few hundred representative turn-switches will be selected for RT experiments with naive, native, Dutch subjects. CGN recordings lack video and were mostly done on Sony-minidisk or otherwise band-limited (for the consequences of using Sony mini-disk recordings in speech research, see Van Son, 2002). To obtain facial/gaze data and wide-band, multichannel audio recordings, around 2 hours of new laboratory dialogs will be recorded on video-tape and audio CD. A selection of these recordings will be annotated, labeled, and segmented (this will be out-sourced to SPEX). Turn switches from this corpus will be added to the existing material.

2c. Innovation
Speech comprehension is an extremely complex phenomenon. To get a grip on it, straightforward, answerable questions are needed. The primary innovation of this project is the selection of a maneagable phenomenon in speech comprehension, Turn Taking in conversations. To be able to select the right time to take a turn, a speaker must have a good understanding of all linguistic aspects of the ongoing turn. However, the actual decision is a simple, yes/no (/maybe) question with minimal demands on the listeners processing capacity. This project asks three simple questions about the identification of turn-relevant places in conversation: What information is available? When is it available? How is it used? The timing information is used to study the underlying processes of *speech comprehension* and *decision making*.
Another innovation is the fact that this project combines off-the-shelf methods in separate areas of (psycho-) linguistics and phonetics to answer these questions. Conventional studies tend to be limited to only one or two levels in speech.
Finally, a new method, laryngography, is recruited to measure the preparation point where the subject (experimental next speaker) realizes that a TRP is imminent and starts preparing for a new utterance. This allows the investigation of early TRP identification processing, something that is difficult with standard RT methods.

2d. Plan of work
Time table

PostDoc (5 year, 0.8 fte)	**AIO (PhD student, 4 year, 1.0 fte)**
2004 months 1-6 - Selection and isolation of relevant material. Inspect 1000 TRP's manually - Video recording of wide-band (clean) laboratory dialoges: Strive for 2 hours of dialog - Start of annotation, labeling, and segmental alignment - Reporting on international meetings and in peer-reviewed literature	
2004 months 7-12 - Continuing of transcription, labeling, and segmentation of material - Preparing transcribed material for listening tests - Start of listening tests (TRP judgments, RT experiments, and laryngography) - Reporting on international meetings and in peer-reviewed literature	
2005 12 months - Completing transcription, labeling, and segmentation of material - Listening tests (using manipulated speech) - Reporting on international meetings and in peer-reviewed literature	**2005 months 1-3** - Familiarizing with the Conversation Analysis (CA) and phonetics literature, and the corpora - Familiarizing with the working environment at the Chair of Phonetics. **months 4-12** - First selection of material to study. - Start of TRP analysis (CA) for selection. - Starting measurements on available material
2006-2008 36 months - Extended listening tests - Selection and analysis of additional speech material to elucidate new/missed aspects of TRP's - Analysis of experiments - Modeling TRP comprehension - Reporting on international meetings and in peer-reviewed literature - Writing final report (2008)	**2006 12 months** - Selection of speech materials - Identifying and analyzing TRP's (CA) of chosen materials - Listening (RT) experiments to determine time course and precision of TRP prediction - Reporting on international meetings and in peer-reviewed literature **2007 12 months** - Listening (RT) experiments to determine time course and precision of TRP prediction - Reporting on international meetings and in peer-reviewed literature **2008 12 months** - Combining and analyzing results - Integrating acoustic and "conventional" aspects of TRP's in a model of TRP prediction. - Writing thesis - Reporting on international meetings and in peer-reviewed literature

Institutional setting

Phonetics research at the University of Amsterdam has long targeted the origins and implications of the variability in speech, going back at least to the work of Koopmans-van Beinum (1980) and Buiting in the seventies. The proposed project would be a natural extension of this long line of research in that it probes even more fundamental questions about the integration of linguistic and acoustic variation in comprehension. The project (and AIO) will benefit from the large knowledge base and contacts of the Chair of Phonetic Sciences on the questions of discourse structure, phonetic variability and managing and handling speech corpora.

We closely collaborate with groups from Helsinki and St. Petersburg in the international *INTAS 915* project, where we study systematic differences in the prosody and acoustics between Russian, Finnish, and Dutch.

The Chair of Phonetic Sciences, in the person of Louis Pols, is already participating in the *CGN*. Within his current project, the applicant has assembled a labeled speech corpus of (monologue) discourses in collaboration with those compiling the *CGN* (Van Son et al., 2001; Van Son and Pols, 2001b&d). Currently we work with the MPI to add this *IFA corpus* to the IMDI/ISLE browsable meta-data corpus ([6]). It is our intention to incorporate speech material from the present project into the *CGN* and *IMDI* corpus whenever possible.

As this project covers a vast range of subjects, I contacted specialists who are willing to help us with advise on specific topics. The construction of a labeled corpus of dialogs requires special expertise that can be found at SPEX. Therefore, we will outsource this part to SPEX (dr. Henk van den Heuvel). Van den Heuvel will advice on the construction of the dialog corpus. Dr. Johanneke Caspers (Leiden University) will advise on the analysis of dialogs and prosody in general. Dr Marc Swerts of Tilburg University (KUB) will advise on video (face) recordings and annotations. Dr Roel Smits from the MPI Nijmegen will advise on listening experiments and be our contact person at the MPI. From the department of Linguistics of the University of Amsterdam, Dr Ingrid C. van Alphen will advise on matters regarding conversation analysis and Dr Henk Zeevat on syntactic parsing and syntactics in general. At the Chair of Phonetics at the University of Amsterdam, there is expertise on the management and evaluation of video-recordings (dr. Jeannette van der Stelt and our technician Ton Wempe), prosodic annotations (dr. Cecilia Odé), and the construction and management of large, on-line corpora (the applicant, dr Rob van Son)

Relevant PhD theses of the Chair of Phonetic Sciences:

Boersma, P.P.G. (1998). Functional Phonology. Formalizing the interactions between articulatory and perceptual drives, Ph.D. thesis, University of Amsterdam, LOT 11, 493p.

Koopmans-van Beinum, F.J. (1980). *Vowel contrast reduction*, PhD thesis, University of Amsterdam, 163p.

Streefkerk, B. (2002). *Prominence: Acoustical and lexical/syntactic correlates*, PhD thesis, University of Amsterdam LOT 58, 152p.

Ten Bosch, L.F.M. (1991). *On the structure of vowel systems: an extended dispersion model*, Ph.D. thesis, University of Amsterdam, 190 p.

Tielen, M.T.J. (1992). *Male and female speech. An experimental study of sex- related voice and pronunciation characteristics*, Ph.D. thesis, University of Amsterdam, 180 p.

Van Alphen, P. (1992). *HMM-based continuous-speech recognition. Systematic evaluation of various system components*, Ph.D. thesis, University of Amsterdam, 216 p.

Van Bergem, D.R. (1995). *Acoustic and lexical vowel reduction*, IFOTT Studies in language and language use 16, Ph.D. thesis, University of Amsterdam, 195 p.

Van der Stelt, J.M. (1993). *Finally a word. A sensori-motor approach of the mother-infant system in its development towards speech*. Ph.D.-thesis, IFOTT Studies in language and language use 4, University of Amsterdam, 226 pp.

Van Donzel, M.E. (1999). *Prosodic aspects of information structure in discourse*, Ph.D. Thesis, University of Amsterdam, LOT 23, 194 p.

Van Son, R.J.J.H. (1993). *see list of publications*

Van Wieringen, A. (1995). *Perceiving dynamic speechlike sounds, Psycho-acoustics and speech perception*, PhD thesis, University of Amsterdam, 256p.

Wang, X. (1997). *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, IFOTT Studies in language and language use 29, Ph.D. thesis University of Amsterdam, 190p.

2e. Literature references (see also under **5. Publications** and **Institutional setting** for our own contributions)

1. Auer, P. (1996). "On the prosody and syntax of turn-continuations", in: Couper-Kuhlen, E. and Selting, M. (eds.) Prosody in conversation: interactional studies, Cambridge University Press, 57-100.
2. Bates, E. and Goodman, J. (1999) "On the Emergence of Grammar from the Lexicon", in B. MacWhinney (ed.), *The Emergence of Language* (Mahwah, NJ: Lawrence Erlbaum Associates), 29-79.
3. Boersma, P. and Weenink, D (1996). *Praat, a System for doing Phonetics by Computer, version 3.4.* Institute of Phonetic Sciences of the University of Amsterdam, Report 132, 182 p.
4. Borsky S., Tuller B. and Shapiro L.P. (1998). "How to milk a coat: The effects of semantic and acoustic information on phoneme categorization", Journal of the Acoustical Society of America 103, 2670-2676.
5. Bradbury, J.W. and Vehrencamp, S.L. (1998). *Principles of animal communication*, Sinauer Associates, 882 p.

6. Broeder, D., Brugman, H. and Wittenburg, P. (2001). "Aspects of modern multi-modal/multi-media corpora exploitation environments", Proc. Eurospeech'2001 Scandinavia, Aalborg, 1529-1531.
7. Byrd, D. and Saltzman, E. (1998). "Intragestural dynamics of multiple prosodic boundaries", Journal of Phonetics 26, 173-199.
8. Caspers, J. (1998). "Who's next? The melodic marking of question versus continuation in Dutch", Language and Speech 41, 375-398.
9. Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. (1997). "The Reliability of a Dialogue Structure Coding Scheme", Computational Linguistics, 23(1), 13-31.
10. Clark, H.H. (2002). "Speaking in time", Speech Communication, 36, 5-13.
11. Cutler A. (1997). The comparative perspective on spoken-language processing, Speech Communication 21, 3-15.
12. De Jong, K., Beckman, M.E., and Edwards, J. (1993). "The interplay between prosodic structure and coarticulation", Language and Speech 36, 197-212.
13. De Jong, K. (1995). "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation", Journal of the Acoustical Society of America 97, 491-504.
14. Eco, U. (1987). Lector in Fabula, Bert Bakker Amsterdam, pp. 333.
15. Fougeron, C. and Keating, P.A. (1997). "Articulatory strengthening at edges of prosodic domains", Journal of the Acoustical Society of America 101, 3728-3740.
16. Fourcin A.J., Abberton E. (1977), "Laryngograph studies of vocal fold vibration", Phonetica 34, 313-315.
17. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). "An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogues", Language and Speech 41, 295-321.
18. Langacker, R.W. (2001). "Discourse in cognitive grammar", Cognitive Linguistics 12-2, 143-188.
19. Massaro, D.W. and Stork, D.G. (1998). "Speech Recognition and Sensory Integration", American Scientist 86, 236-244.
20. McGurk H. and MacDonald J. (1976). "Hearing lips and seeing voices" Nature, Dec. 1976, 746-748.
21. Norris D., McQueen J.M., and Cutler A. (2000). "Merging information in speech recognition: Feedback is never necessary", Behavioral and Brain Sciences 23, 299-325.
22. Norris, D. (2001). "Modelling speed and accuracy in behavioural experiments on speech recognition", Proceedings of the workshop on speech recognition as pattern classification, Nijmegen, 11-13th July, 79-84.
23. Oostdijk, N. (2000). "The Spoken Dutch Corpus, overview and first evaluation", Proceedings of LREC-2000, Athens, Vol. 2, 887-894.
24. Pettersson, J.S. (1996). *Grammatological Studies: Writing and its Relation to Speech* PhD thesis, Department of Linguistics, Uppsala University, Published as Reports from Uppsala University Linguistics no. 29, pp.228.
25. Pickering, M.J. and Branigan, H.P. (1999). "Syntactic priming in language production", Trends in Cognitive Sciences 3, 136-140.
26. Pinker, S. (1999). Words and rules: The ingredients of language, Perseus books, 352p.
27. Sacks, H., Schegloff, E.A., and Jefferson, G. (1974). "A simplest systematics for the organization of turn-taking for conversation", Language 50, 696-735.
28. Schegloff, E.A. (1999). "Discourse, pragmatics, conversation, analysis", Discourse Studies 1(4), 405-435.
29. Schegloff, E.A (2000). "Overlapping talk and the organization of turn-taking for conversation", Language in Society 29, 1-63.
30. Shimojima, A., Katagiri, Y., Koiso, H. and Swerts, M. (2002). "Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses", Speech Communication, 36, 113-132.
31. Smith, W.J. (1977). The behavior of communicating, Harvard University Press, Cambridge Massachusetts and London, 545p.
32. Stent, A. (2002). "A conversation acts model for generating spoken dialogue contributions", Computer Speech and Language 16, 313-352.
33. Tanaka, H. (2000). "The particle ne as a turn-management device in Japanese conversation", Journal of Pragmatics 32, 1135-1176.
34. Thompson, H.H. (1996). "Why 'Turn-taking' is the wrong way to analyze dialogue: Empirical and theoretical flaws", Proceedings of the ISSD, Philadelphia, paper B.4.1, 49-52.
35. Turk, A.E. and Sawush, J.S. (1997). "The domain of accentual lengthening in American English", Journal of Phonetics 25, 25-41.
36. Wells, B. and Peppé, S. (1996). "Ending up in Ulster: prosody and turn-taking in English dialects", in: Couper-Kuhlen, E. and Selting, M. (eds.) Prosody in conversation: interactional studies, Cambridge University Press, 101-130.
37. Ward, N. (1996). "Using prosodic clues to decide when to produce back-channel utterances", Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia (ICSLP-96), 1728-1731.
38. Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P.J. (1992). "Segmental durations in the vicinity of prosodic phrase boundaries", Journal of the Acoustical Society of America 91, 1707-1717.

[**2f. Utilisation paragraph**: Only required for proposals to be submitted to Technical Sciences, see Notes

Vernieuwingsimpuls / Innovational Research

Grant application form 2003 **VIDI scheme**

Please refer to Explanatory Notes when completing this form

Cost estimates

3a. Budget

	200y	200y+1	200y+2	200y+3	200y+4	TOTAL
Staff costs: (in k€)						
Applicant 0.8 fte	12	12	12	12	12	350
Post-doc						
Ph D student (AIO) 1.0 fte		12	12	12	12	110
Support staff						
Non staff costs: (k€)						
Equipment						
Consumables	30	30	12	4	4	80
Travel and subsistence	5	13.5	10	17	10	55.5
Other						
TOTAL						595.5

Travel costs include conference participation and visits to laboratories abroad. In this field there is at least one large conference each year (e.g., EUROSPEECH or ICSLP) and numerous workshops. For each participant, visits to one conference and one workshop a year are planned (2x Euro 2500,- a year, on average). Also an extended visit to another laboratory is planned for the graduate student (Euro 3500,-) and two for the post-doc (together Euro 7000,-). Good options would be the HRCR at Edinburgh University, the MPI at Nijmegen, and the University of Freiburg (Prof. Auer).

Material costs cover speakers and listeners (informants), recording and backup media, labeling, and segmentation of the speech. We expect Euro 9,000 for “informants”, Euro 1,000 for media, and Euro 70,000 for construction of the corpus. Most of this money will be spent in the first two years for the construction of the annotated video corpus. In the remaining years money will still be needed for informants and recording media.

The labeled corpus will be constructed according to the standards of the *CGN* to allow future integration. Copyrights for all original material will be transferred to the *Nederlandse Taalunie* (Dutch Language Union) as was done with the copyrights of the *CGN* and the *IFA corpus*. These materials will be released under a Free Software license (the *GNU General Public License*, or *GPL* for short). These materials will also be added to the IMDI/ISLE browsable meta-data corpus. We will try to obtain permission to re-distribute any material obtained from others under the GPL or another open source license, e.g., *CGN* recordings. The GPL ensures the access and availability of the corpus beyond the duration of this project. From our work with the *IFA corpus* we have learned that this is a highly efficient arrangement.

3b. Have you requested any additional grants for this project either from NWO or from any other institution? no (If ‘yes’, see Notes)

Curriculum vitae

4f. Brief summary of research over last five years

(max. 350 words)

The last five years of research spans four distinct lines of research. The first line studied the acoustics of consonant reduction. Consonants showed acoustic differences that completely parallel those found in vowels when reduction is involved. Acoustic reduction in consonants was most evident from duration and spectral balance measurements. Perceptual experiments confirmed a link between acoustic measures of reduction and reduced intelligibility of consonants (Van Son and Pols, 1997, 1999a).

A second line of research studied the time-course in which listeners extract information from speech for phoneme identification. Listeners use peri-segmental (extra-segmental) speech to identify phonemes from connected speech. It was concluded that listeners use all information available, whatever its origin, to decide as quickly as possible on phoneme identity (Van Son and Pols, 1999b, 2001a&c). Also, phoneme identification in the classical experimental paradigms can best be explained as an automatic, non-exclusive, categorization, followed by access into the mental lexicon (Van Son and Pols, 2001 a&c).

The third line of research investigates the way speech is made efficient by varying the "effort" put in producing its acoustic signal with communicative importance. For this research, a 50 kWord, phonemically segmented database of speech has been constructed (Van Son et al, 2001; Van Son and Pols, 2001 b&d; Pols and Van Son, 2002; Van Son and Pols, 2002; Van Son and Pols, accepted). At the upcoming ICPhS 2003 conference at Barcelona we will present results which show a consistent correlation between the importance of a vowel or consonant for word recognition in context and its level of reduction. Other papers are in preparation. This line of research also includes the participation in the international INTAS 915 project, a collaboration between groups in St. Petersburg, Helsinki, and the Chair of Phonetic Sciences in Amsterdam, which studies the language specific differences in prosodic and acoustic variation between Russian, Finnish, and Dutch.

A fourth, ongoing branch of research is the construction and maintenance of on-line resources, most notably, the 50 kWord IFA speech corpus (Van Son et al, 2001; Van Son and Pols, 2001 b&d; Van Son, 2002).

4g. International activities

4h.Other academic activities

Teaching:

Vernieuwingsimpuls / Innovational Research

Grant application form 2003 **VIDI scheme**

Please refer to Explanatory Notes when completing this form

List of publications

5. Publications:

(As the proposed research is really new, there are no publications that completely overlap with the intended work. Instead, the publications that were instrumental in designing this proposal, both for the theoretical and practical aspects, are marked with an **S**.)

-International (refereed) journals

Van Son, R.J.J.H. and Pols, L.C.W. (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", J. Acoust. Soc. Am. 88(4), 1683-1693.

Van Son, R.J.J.H. and Pols, L.C.W. (1992). "Formant movements of Dutch vowels in a text, read at normal and fast rate", J. Acoust. Soc. Am. 92, 121-127.

S Van Son, R.J.J.H. and Pols, L.C.W. (1999a). "An acoustic description of consonant reduction", Speech Communication 28, 125-140.

S Van Son, R.J.J.H. and Pols, L.C.W. (1999b). "Perisegmental speech improves consonant and vowel identification", Speech Communication, 29, 1-22.

S Van Son, R.J.J.H, Binnenpoorte, D., van den Heuvel, H. and Pols, L.C.W. (2001). "The IFA Corpus: A phonemically segmented Dutch "Open Source" speech Database" Proceedings of Eurospeech 2001, Aalborg, Denmark, 811-814.

S Van Son, R.J.J.H. and Pols, L.C.W. (2002). "Evidence for Efficiency in vowel production", Proceedings of ICSLP2002, Denver, USA, .

S Pols, L.C.W. and Son, van R.J.J.H. (1993). "Acoustics and perception of dynamic vowel segments", Speech Communication, 13, 135-147.

Van der Heijden, L.A.M. and Van Son, R.J.J.H. (1982). "A note on the neglect of the doppler effect in the modelling of traffic flow as a line of stationary point sources", Journal of Sound and Vibration 85, 442-444. (this paper was the result of an undergraduate project, both authors contributed equally to it)

Van Son R.J.J.H. and Van Santen, J.P.H. (final phase of review). "Duration and spectral balance of intervocalic consonants", submitted to the Journal of the Acoustic Society of America.

-National (refereed) journals

-Books, or contributions to books

Van Son, R.J.J.H. and Pols, L.C.W. (1993). "How does speaking rate influence vowel formant track parameters", In: V.J. van Heuven and L. C.W. Pols (Eds.), *Analysis and synthesis of speech. Strategic research towards high-quality text-to-speech generation*, Mouton de Gruyter, Berlin, 171-191.

S Van Son, R.J.J.H. (1993). S*pectro-temporal features of vowel segments*, Ph.D. thesis, IFOTT Studies in language and language use 3, University of Amsterdam, 195p.

Pols, L.C.W. and Van Son, R.J.J.H. (1996). "Acoustics and perception of dynamic vowel segments", in *Analysis, perception and processing of spoken language*, G. Fant, K. Hirose, S. Kiritani (eds.), Elsevier, North-Holland, 135-147.

Pols, L.C.W. and Van Son, R.J.J.H. (2002). "Accessing the IFA-corpus", In: *Book in honor of the 70-th anniversary of Prof. L.V. Bondarko*, N.B. Volskaya, N.D. Svetozarova (Eds.), University of St. Petersburg, 316-320.

S Pols, L.C.W. and Van Son, R.J.J.H. (accepted for publication). "Speech Dynamics: Acoustic manifestations and perceptual consequences", in *Dynamics of Speech Prodution and Perception*, Pierre Divenyi, Georg Meyer and Klara Vicsi (eds.), IOS Press, Amsterdam, Washington, Oxford.

-Other

Corpora

S Van Son, R.J.J.H, Binnenpoorte, D., van den Heuvel, H. and Pols, L.C.W. (1999-2003). "The IFA corpus", a 50 kWord fully labeled and segmented corpus (2003). Published as an IMDI/ISLE browsable corpus at the Max Planck Institute for Psycholingistics, Nijmegen (IMDI browser: http://www.mpi.nl/world/corpora/IFAcorpus/IMDI/IFAcorpus.imdi)

Conference and Institutional proceedings

Van Son, R.J.J.H. (1987). "Automatic slope measurements on formant tracks", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 11, 67-78.

Van Son, R.J.J.H. (1988a). "Automatische segmentatie en stilering van parameterkrommen", Proceedings Colloquium Signaalanalyse en Spraak, Leidschendam.

Van Son, R.J.J.H. (1988b). "Differences in formant values of Dutch vowels due to speaking rate", Proceedings of Speech 88, 7th FASE symposium, Edinburgh, 313-320.

Van Son, R.J.J.H. and Pols, L.C.W. (1989). "Comparing formant movements in fast and normal rate speech", Proceedings Eurospeech 89, Paris, Vol. 2, 665-668.

Van Son, R.J.J.H. and Pols, L.C.W. (1991a). "The influence of formant track shape on the perception of synthetic vowels", Proceedings Eurospeech 91, Genova, Vol. 3, 1117-1120.

Van Son, R.J.J.H. and Pols, L.C.W. (1991b). "The influence of speaking rate on vowel formant track shape as modeled by Legendre polynomials", Proceedings of the Institute of Phonetic Sciences Amsterdam 15, 43-59.

Van Son, R.J.J.H. and Pols, L.C.W. (1993). "Vowel identification as influenced by vowel duration and formant track shape", Proceedings Eurospeech 93, Berlin, 285-288.

Van Son, R.J.J.H. (1993). "Vowel perception: a closer look at the literature", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 17, 33-64.

Van Son, R.J.J.H. (1994). "A method to quantify the error distribution in confusion matrices", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 18, 41-63.

Van Son, R.J.J.H.(1995a). "A method to quantify the error distribution in confusion matrices", Proceedings Eurospeech 95, Madrid, 2277-2280.

Van Son, R.J.J.H. (1995b). "The relation between the error distribution and the error rate in identification experiments", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 19, 71-82.

Van Son, R.J.J.H. and Pols, L.C.W. (1995a). "The influence of local context on the identification of vowels and consonants", Proceedings Eurospeech 95, Madrid, 967-970.

Van Son, R.J.J.H. and Pols, L.C.W. (1995b). "What does consonant reduction look like, if it exists", Proceedings Eurospeech 95, Madrid, 1909-1912.

Van Son, R.J.J.H. and Pols, L.C.W. (1995c). "How transitions and local context affect segment identification", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 19, 51-69.

Van Son, R.J.J.H. and Pols, L.C.W. (1995d). "Acoustic consonant reduction: A comparison", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 19, 83-91.

Van Son, R.J.J.H. and Pols, L.C.W. (1996a). "An acoustic profile of consonant reduction", Proceedings of ICSLP 96, Philadelphia, USA, 1529-1532.

Van Son, R.J.J.H. and Pols, L.C.W. (1996b). " A comparison between the acoustics of vowel and consonant reduction ", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 20, 13-25.

S Van Son, R.J.J.H. and Pols, L.C.W. (1997). "The correlation between consonant identification and the amount of acoustic consonant reduction", Proceedings Eurospeech 97, Rhodes, 2135-2138.

Van Son, R.J.J.H. and Van Santen, J.P.H. (1997). "Strong interaction between factors influencing consonant duration", Proceedings Eurospeech 97, Rhodes, 319-322.

S Van Son, R.J.J.H.., Koopmans-van Beinum, F.J. and Pols, L.C.W. (1998). "Efficiency as an organizing principle of natural speech", Proceedings ICSLP 98, Sydney, Australia, 2395-2398.

S Van Son, R.J.J.H and Pols, L.C.W. (1999). "Effects of stress and lexical structure on speech efficiency", Proceedings of Eurospeech 99, Budapest, Hungary, 439-442.

S Van Son, R.J.J.H, Streefkerk, B.M. and Pols, L.C.W. (2000). "An acoustic profile of speech efficiency", Proceedings of ICSLP2000, Beijing,, China, IV 97-100.

S Van Son, R.J.J.H and Pols, L.C.W. (2001a). "Phoneme recognition as a function of task and context", Proceedings of the SPRAAC workshop, Nijmegen, the Netherlands, 25-30.

S Van Son, R.J.J.H and Pols, L.C.W. (2001b). "The IFA Corpus: A phonemically segmented Dutch "Open Source" speech Database", Proceedings of the IRCS workshop on Linguistic Databases, Philadelphia, 245-253.

S Van Son, R.J.J.H and Pols, L.C.W. (2001c). "Phoneme recognition as a function of task and context", Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 24, 27-38.

S Van Son, R.J.J.H and Pols, L.C.W. (2001d). "The IFA Corpus: A phonemically segmented Dutch "Open Source" speech Database" Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 24, 15-26.

Van Son, R.J.J.H. (2002). “Can standard analysis tools be used on decompressed speech?”, Paper presented at the COCOSDA2002 meeting, Denver (URL:http://www.cocosda.org/meet/denver/COCOSDA2002-Rob.pdf).

Van Son, R.J.J.H and Pols, L.C.W. (accepted for publication). "An acoustic model of communicative efficiency in consonants and vowels taking into account context distinctiveness", Proceedings of the Interactional Conference of Phonetic Sciences, Barcelona.

Reports

Van Son, R.J.J.H. (1986). "Lawaai en dieren", Wetenschapswinkel van de Universiteit van Amsterdam.

Van Son, R.J.J.H. and Pols, L.C.W. (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", SPIN-ASSP report 16, Stichting Spraaktechnologie, Utrecht, 22p.

Van Son, R.J.J.H. and Pols, L.C.W. (1991). "Aspects of speech rate on spectro-temporal characteristics of speech", Final report SPECRED, SPIN-ASSP report 42, 49p.

Van Son, R.J.J.H. (1992). "FORM: documentation", Report of the Institute of Phonetic Sciences Amsterdam 118, 87p.

Van Son, R.J.J.H. (1994). "Alfeios, a multi stream rule compiler" Manual for the Alfeios rule development package developed for the ESPRIT POLYGLOT program, Department of Language and Speech, Nijmegen University, 58p.

Signature

I hereby declare that I have completed this form truthfully:

Please submit the application to NWO in electronic form (pdf format is required!) using the IRIS system, which can be accessed via the NWO website (www.nwo.nl/vernieuwingsimpuls).

Vernieuwingsimpuls / Innovational Research

Grant application form 2003

VIDI scheme

Please refer to Explanatory Notes when completing this form

Post to NWO / Vernieuwingsimpuls

To streamline the processing of applications, please complete the form below and post a print-out of this page together with any relevant documents to NWO.

I the undersigned declare that I have today posted (tick relevant documents):

Institutional guarantee from Board ('Inbeddingsgarantie College van Bestuur')
The board will contact NWO directly.

Address list of 'non-referees'
Not needed

Name of applicant: Rob van Son

Place: Amsterdam

Date: 13 Februari 2003

Postal address: Chair of phonetic sciences
University of Amsterdam
Herengracht 338
1016 CG Amsterdam
The Netherlands

NWO Council area: GW

Send the documents to:

NWO/Vernieuwingsimpuls
Council area: GW
P.O. Box 93138
2509 AC The Hague
(The Netherlands)