2010-01-01

A Study of Accomodation of Prosodic and Temporal Features in Spoken Dialogues in View of Speech Technology Applications

Spyridon Kousidis [Thesis]
Technological University Dublin

A Study of Accommodation of Prosodic and Temporal Features in Spoken Dialogues in View of Speech Technology Applications

PhD Thesis

2010

Spyridon Kousidis

Digital Media Center
School of Media
Dublin Institute of Technology

Research Supervisors: **Dr. David Dorran**

Prof. Ciaran MacDonnaill

Prof. Eugene Coyle

I certify that this thesis which I now submit for examination of the award of PhD is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Dublin Institute of Technology and has not been submitted in whole or in part for an award in any other Institute or University.

The work reported on in this thesis conforms to the principles and requirements of the Institute for ethics in research work.

The Institute has permission to keep, to lend or to copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature

Date

Abstract

Inter-speaker accommodation is a well-known property of human speech and human interaction in general. Broadly it refers to the behavioural patterns of two (or more) interactants and the effect of the (verbal and non-verbal) behaviour of each to that of the other(s). Implementation of this behavior in spoken dialogue systems is desirable as an improvement on the naturalness of human-machine interaction. However, traditional qualitative descriptions of accommodation phenomena do not provide sufficient information for such an implementation. Therefore, a quantitative description of inter-speaker accommodation is required.

This thesis proposes a methodology of monitoring accommodation during a human or human-computer dialogue, which utilizes a moving average filter over sequential frames for each speaker. These frames are time-aligned across the speakers, hence the name Time Aligned Moving Average (TAMA). Analysis of spontaneous human dialogue recordings by means of the TAMA methodology reveals ubiquitous accommodation of prosodic features (pitch, intensity and speech rate) across interlocutors, and allows for statistical (time series) modeling of the behaviour, in a way which is meaningful for implementation in spoken dialogue system (SDS) environments.

In addition, a novel dialogue representation is proposed that provides an additional point of view to that of TAMA in monitoring accommodation of temporal features (inter-speaker pause length and overlap frequency). This representation is a percentage turn distribution of individual speaker contributions in a dialogue frame which circumvents strict attribution of speaker-turns, by considering both interlocutors as synchronously active. Both TAMA and turn distribution metrics indicate that correlation of average pause length and overlap frequency between speakers can be attributed to accommodation (a debated issue), and point to possible improvements in SDS "turn-taking" behaviour.

Although the findings of the prosodic and temporal analyses can directly inform SDS implementations, further work is required in order to describe inter-speaker accommodation sufficiently, as well as to develop an adequate testing platform for evaluating the magnitude of perceived improvement in human-machine interaction. Therefore, this thesis constitutes a first step towards a convincingly useful implementation of accommodation in spoken dialogue systems.

This thesis is dedicated to my father, Dr Ilias Kousidis (1940-2006)

Acknowledgments

The work presented in this dissertation would not have been possible without the support and encouragement of a number of individuals. I would like to express my sincere gratitude to the following persons:

Dr David Dorran, for his invaluable support and encouragement, as well as his dedication to perfection when providing feedback throughout the 4-year course of this project.

Prof. Ciaran MacDonnaill, for sharing his experience and insight in conducting research and following through the set objectives and goals. Also for being there at difficult times.

Prof. Eugene Coyle, for believing in my ability to undertake academic research work since our first acquaintance in the School of Electrical Engineering some years ago. Also for his always “right-on-the-spot” advice.

Dr Charlie Cullen, Dermot Campbell, Brian Vaughan and Yi Wang, for their great team-work spirit and collaboration during the SALERO project. It was enjoyable working with you, and I hope we may work together again in the future. Also to everyone in the DMC, for helping out in numerous occasions and participating in the recording experiments.

Charlie Pritchard, managing director of the Digital Media Center, for always doing his best in providing for all requirements throughout this research project. Also for his caring attitude towards individual researchers' problems and assisting in any way he can.

Simon Bursell, for implementing the code modifications to the FreeTTS player, which was used in the SDS environment testing platform.

Special thanks to Mietta Lennes, Prof. Plinio Barbosa, Prof. Hugo Quené, Ninva H. de Jong and Ton Wempe, for making their Praat scripts available. Also to Prof. Nick Campbell, for making annotations of Japanese telephone dialogues publicly available.

Jens Edlund, Prof. Joakim Gustafson, Prof. Mattias Heldner and Prof. Stephan Benus for their comments and positive feedback on my work.

To my family, who have always been there for me, and to my close friends, for constantly encouraging me to “go on”.

Finally, to Despoina, for making my life happy, showing patience and understanding, as well as helping out with dialogue annotations.

Table of Contents

Table Index

Figure Index

Equation Index

List of abbreviations

a/p:	*Acoustic/prosodic*
ALU:	*Automatic Language Understanding*
ANOVA:	*ANalysis-Of-VAriance*
APL:	*Average Pause Length*
AR(n):	*Auto-Regressive model of order n*
ASR:	*Automatic Speech Recognition*
CAT:	*Communication Accommodation Theory*
CP & PP:	*Comprehension – production and production – production (priming)*
ER:	*Exchange Rate*
F0:	*Fundamental frequency*
IAM:	*Interactive Alignment Model*
IAT:	*Interpersonal Adaptation Theory*
IVR:	*Interactive Voice Response (systems)*
JAT:	*Joint Active Time*
MIP:	*Mood Induction Procedure(s)*
MOS:	*Mean Opinion Score (tests)*
MSE	*Mean Square Error*
NLP:	*Natural Language Processing*
OR:	*Overlap Rate*
SAT:	*Speech Accommodation Theory*
SDS:	*Spoken Dialogue System(s)*
SPL:	*Sound Pressure Level*
TAMA:	*Time-Aligned Moving Average*
TO:	*Total Overlap (duration)*
TRP:	*Transition Relevant Points*
TS:	*Turn Share*
TTS:	*Text-To-Speech (synthesis)*
VAD:	*Voice Activation Detection*
VAR (n):	*Vector Auto-Regressive model of order n*
VARX:	*VAR model including eXogenous factors*

1 Introduction

1.1 Overview

The phenomenon of inter-speaker accommodation in spoken dialogues is well-known in psycholinguistics, communication and cognitive sciences. The term itself is one of many used to describe a variety of complex phenomena associated with two – or more – interlocutors and the tendency of various features of their verbal and non-verbal behaviour to display growing similarity over time as the dialogue evolves, or across several dialogue sessions. The features span the entire spectrum of forms of human expression: lexical, syntactic, prosodic, gestural and postural features, as well as turn-taking behaviour have been found to converge across interlocutors engaging in dialogue, both in controlled laboratory experiments, as well as in naturally occurring conversations.

The utilization of such behaviour in speech technology applications is highly desirable, for a variety of reasons. First, it opens an avenue of improvement upon the naturalness of synthesized speech, in the context of spoken dialogue systems (SDS), as it may be possible for the system voice to adapt to that of the user, providing for a more pleasant conversation. Second, accounting for accommodation can improve the overall performance of on-line monitoring, a process which is vital in predicting user expectations and user satisfaction/frustration in real time. Finally, implementation of temporal accommodation is essential in establishing a more sophisticated interaction management strategy in SDS applications, for the purpose of providing smoother and more efficient human-machine interaction.

However, direct implementation of inter-speaker accommodation into current speech technology applications is not feasible for two reasons: first, inter-speaker accommodation is a complex behavioural phenomenon, and its manifestation in spoken language has not been quantitatively described yet; and second, current SDS architectures are not designed to accommodate natural dialogue with human users, therefore a platform for testing quantitative models of inter-speaker accommodation does not yet exist. This thesis focuses on the first problem, i.e. the quantitative description of accommodation phenomena in view of implementation in spoken dialogue interfaces and systems, but also presents a preliminary application of the accommodation models in a simulated SDS environment.

1.2 Scope of work and motivation

This thesis focuses on the description of accommodation phenomena exhibited in specific properties of speech. In particular, the features studied are acoustic-prosodic measures of the speech signal (pitch, intensity and speech rate), as well as temporal features (inter-speaker silence duration and occurrence of overlapping speech). The motivation for studying these specific features is explained

in detail in chapter 2. Briefly, the acoustic-prosodic (a/p) features were selected because of their historical prominence in improvements on naturalness of *synthesized* speech, which in turn is due to the multitude of functions they are known to carry in *human* speech: prosody is essential in speech production and perception (Cutler *et al.* 1997); expresses attitudes and emotions (Schroeder *et al.* 2001); and enables smooth dialogue transitions and non-interrupting overlapping speech which provides feedback to the speaker (Cerrato 2002). Temporal features are also central to the temporal organization of dialogue, i.e. the smooth transition of turns between interlocutors. Importantly, the function of turn-taking and smoothness of dialogue is one of the major short-comings in current SDS applications (Raux and Eskenazi 2008).

A better understanding of the accommodation phenomena related to prosodic and temporal features may directly improve the performance of current SDS technology in various ways: (a) enhancement of the human metaphor (Edlund *et al.* 2008), by simulating accommodation in SDS, (b) smoothness of conversational dialogue based on temporal accommodation, (c) positive evaluation of the overall interaction by the user, based on convergence, according to certain theories (e.g. Giles *et al.* 1987), (d) improvement of prosodic models for synthesized speech, in relation to the problem of mapping abstract prosodic representations to actual signal features, by providing more appropriate prosodic baselines, (e) informing classification for emotion recognition in dialogues (Cowie *et al.* 2001), (f) informing classification of dialogue acts (Wright 1999), and (g) improving performance of ASR by exploiting user adaptation to the system voice (Bell *et al.* 2003).

1.3 Aims and objectives

The overall aim of this research was to study accommodation of prosodic and temporal features of speech in human dialogues, in order to inform implementation of this behaviour in spoken dialogue systems. However, as was mentioned in section 1.1, neither a complete theoretical description of accommodation phenomena nor a standard development and testing platform currently exist in order to "build" such a system. Thus a set of more realistic objectives were defined, which are consistent with currently adopted methodologies in speech technology and speech science research in general, and SDS in particular:

a) Design and implementation of a recording laboratory environment, properly equipped in terms of audio equipment and other infrastructure, for the purpose of carrying out recording experiments, specifically for the purpose of acquiring recordings of spontaneous speech (dialogues).

b) Development of tools for annotation and feature extraction of spontaneous speech corpora,

for the purpose of statistical analysis of prosodic and temporal features.

c) Formulation of a methodology for analysis of prosodic features from the recorded dialogues in (a) above, for the purpose of analyzing inter-speaker accommodation within single interactions.

d) Formulation of a methodology for the purpose of analyzing temporal accommodation in human dialogues.

e) Formulation of a quantitative model of accommodation of prosodic and temporal features for implementation in human-computer dialogues

f) The development of a testing platform, in order to demonstrate the implementation of accommodating behaviour in an SDS application environment.

1.4 Contributions

The major contributions presented in this thesis are as follows:

(a) A methodology for monitoring accommodation in human (or human-machine) dialogues (Time-Aligned Moving Average or TAMA in short), that is feature independent and uses summary statistics (average and standard deviation) of normalized speech features in overlapping dialogue frames. The transformation allows direct comparison of features from two speakers in contemporaneous frames.

(b) Statistical evaluation of accommodation of a/p features among speakers in the recorded dialogues by means of time series analysis which verifies and objectively measures the presence of *feedback* and *bi-directional* accommodation. In addition, the statistical methodology points towards possible implementations of similar behaviour in SDS, using models derived from the human dialogues.

(c) A novel dialogue representation (turn-share and turn-distribution) that is complementary to current *chronographic* (Lennes and Anttila 2002) representations of the temporal structure of dialogue (turn-taking). This representation provides evidence additional evidence of temporal accommodation to that previously available, and points to design strategies that can be utilized in SDS implementations.

In addition, the following minor contributions are also presented:

(d) The acquisition of a corpus of spontaneous dialogues, recorded at high audio quality (192 KHz/24-bit) in low-noise conditions (isolation soundproof booths). Each speaker has

been recorded in a separate audio channel, thus eliminating cross-channel noise contamination. Further, each speaker's speech stream has been annotated for silence/vocalization, facilitating future research on this data. The recordings took place in an audio recording laboratory that was setup specifically for the purpose of recording spontaneous speech. The recording experiments and the laboratory setup were collaborative work undertaken within the SALERO project[1].

(e) An exploratory application of the findings from chapter 7 was carried out as a simulated SDS environment with conversational capabilities that adapted its a/p features in accordance to those of human subjects. Despite the fact that this was an experimental approach, thus not performance-optimized, a number of useful conclusions were drawn that could serve in designing actual systems in the future.

1.5 Thesis statement

Human dialogues exhibit accommodation of a/p (pitch, pitch range, speech rate, intensity) and temporal (pause duration and overlaps) features. This thesis proposes quantitative descriptions of these phenomena that provide useful insights for the development of SDS which are capable of implementing appropriately similar behaviour.

1.6 Outline

The following is a brief outline of the rest of this document. Chapter 2 provides relevant background on significant improvements towards natural speech interaction, and naturalness in speech technology in general. Emphasis is placed on prosody, which is central to improving naturalness in various areas of speech technology research. In addition, the major issues in current research and commercial SDS are discussed, in order to highlight areas in which a quantitative description of inter-speaker accommodation can improve on the current performance. Chapter 3 outlines the theoretical frameworks of inter-speaker accommodation, with a more detailed description of Communication Accommodation Theory and the Interactive Alignment Model, which have been the most prominent theoretical descriptions of inter-speaker accommodation in speech technology literature. Chapter 4 provides a review of related work, which consists of previously proposed methods of measuring inter-speaker accommodation. This review is not restricted to prosodic and temporal features but includes studies focusing on accommodation in other features, such as lexical or gestural. A summary and analysis of the literature review findings is presented in chapter 5. The result of this critical analysis provides justification for the rest of the

1 www.salero.info

work presented in this thesis.

Chapter 6 presents the outline of the research methodology followed, as well as the steps taken towards acquiring the data on which the major contributions were based, which comprise the design of the audio recording laboratory and recording experiments for acquiring recordings of spontaneous speech, as well as the development of corpus annotation and feature extraction tools. The TAMA analysis method is presented in chapter 7, along with the statistical evaluation of accommodation of a/p features in the recorded dialogues, which points to possible models for accommodation that can be used in spoken dialogue systems. Chapter 8 presents an application of the TAMA methodology on temporal features, which shows partial evidence of accommodation of pause length and overlap frequency in the recorded dialogues. The novel dialogue representation, also presented in this chapter, explores a different approach to explaining the variations of these features as a function of dialogue activity (or liveliness) and turn share distribution among speakers. The preliminary application of a/p feature accommodation in a simulated SDS environment is presented in chapter 9, and chapter 10 presents the conclusions derived and possible paths for extension of this work.

2 Background

2.1 Overview

This chapter serves as background, in order to explain the motivation behind studying accommodation of acoustic/prosodic (a/p) and temporal features in human dialogues. Since this research is focused on studying these phenomena in view of incorporating them in human-machine interaction, a review of current methodologies employed in spoken interface applications - and their limitations - is essential.

In particular, section 2.2 discusses current research in speech as an interface in human-machine interaction and the need for more "natural", or "human-like" interaction with "talking" systems. As was mentioned in the introduction, this has been the major motivation for the work described in this thesis, as inter-speaker accommodation is a well-known property of human dialogues that could improve the perceived naturalness of human-machine interaction. The issue of naturalness and its evaluation in spoken dialogue systems is also discussed in this section, and the need for corpora of natural human dialogues in order to study inter-speaker accommodation is identified.

Section 2.3 presents a conceptual view of the operation of spoken dialogue systems, indicating some of their advances and limitations that are related to the naturalness of the interaction, as experienced by the users of such systems. The floor-taking and floor-releasing (in short, turn-taking) strategy of SDS is the most notable such limitation, and inter-speaker accommodation of temporal features (such as inter-speaker pause length) is closely related to this problem. Other functions of SDS, such as user monitoring and error detection, are likely to benefit from a quantitative description of prosodic accommodation, as prosody has been used prominently as a classifier in error detection. These findings have motivated the study of inter-speaker accommodation in temporal and a/p features, respectively.

Speech prosody is discussed in detail in section 2.4, in relation to its form (the speech signal features that are considered in the study of prosody) and function (the role of prosody in human speech). Historically, prosody has been prominent in speech technology in relation to the naturalness of synthesized speech as it carries several linguistic and paralinguistic functions. The latter include the expression of emotions as well as prosodic cues that serve dialogue organization, which find application in SDS that detect user emotions (e.g. for error detection) or adapt their turn-taking strategy based on online prosodic analysis of the user utterances. In addition, SDS rely on models of prosody when generating their prompts to the user, with the majority of these models being based on monologue speech. Thus, SDS do not take into account the interaction context when generating prompts, which can make the latter sound inappropriate. A description of inter-speaker

accommodation of a/p features is likely to improve performance in all these areas (prompt generation, emotion recognition/synthesis, online prosodic analysis for interaction management), thus a/p features are identified as primary targets for such a description.

Finally, section 2.5 briefly discusses the issue of acquiring recorded corpora of natural human speech, which is an essential step in studying any property of human speech. This also true for inter-speaker accommodation, in which case recordings of dialogues are required. In particular possible methods of acquiring such recordings in a laboratory environment are compared against using existing data (e.g. from customer service call-centers) according to three specific criteria: naturalness of the content, audio quality and re-usability/resource cost. This chapter is concluded by a discussion that summarizes the key points in section 2.6.

2.2 Towards natural spoken dialogue interfaces

This section discusses current research directions that point towards the development of spoken dialogue systems capable of engaging in "natural" interaction with human users. In particular, sections 2.2.1 and 2.2.2 discuss the motivation and aims of this research direction, as the work presented in this thesis supports the development of these goals. The issue of naturalness is discussed in section 2.2.3, while a framework for implementation and evaluation of human-like behaviour in spoken dialogue systems is discussed in section 2.2.4.

2.2.1 Long term goals in speech science and technology research

A speech-based interface utilizes speech as input and/or output, in order to accomplish an application related task. From this point of view, traditional speech-related technologies, such as text-to-speech synthesis (TTS), automatic speech recognition (ASR), and spoken dialogue systems (SDS), can all be seen as speech interfaces: TTS screen readers are programs that "read aloud" the contents of the computer screen, thus replacing the need for a human reader (Dutoit 1997); typical applications of ASR are "dictation" of text to the computer (replacing the need to type) and voice commands (Boves and Os 1999). Both TTS and ASR have general applications of this type but have also been specifically targeted for people with hearing and/or visual impairments (Syrdal *et al.* 1994; Dutoit 1997; Sproat *et al.* 1999). Spoken dialogue systems, which build upon the other two technologies, are interactive conversational environments with multiple applications, such as automated customer service (Hardy *et al.* 2004), travel booking (Seneff and Polifroni 2000), or call routing (Williams and Witt 2004).

For each of the above applications, speech technology components, such as TTS, ASR and SDS,

aim to *adequately imitate* the human abilities of *speaking, perceiving speech,* and *engaging in conversation.* There are at least two questions arising from the above statement: first, why do humans need "machines" that can speak, understand speech, and engage in conversation? and, secondly, how feasible is this, or, how *adequately* can these machines imitate humans?

The first question can be answered – both from a research and commercial point of view – by the need for speech technology in applications such as those mentioned above. However, there is a much wider scope in this field of research than that which is revealed by the applications themselves. Generally, human scientists and engineers aim to imitate nature and this applies also to speech[2], which is the most natural form of human communication (Lustgarten and Juang 2003); hence the science-fiction incarnations of intelligent androids, or - in the case of an intelligent talking computer – the famous HAL from the film *2001:A space Odyssey*[3]. Although current state-of-the-art systems produce highly intelligible speech, as well as impressive "understanding" capabilities, HAL continues to remain in the sphere of science fiction (Larsson 2005). This also answers the second question: how good are the current systems? In short, not good enough, according to (Pieraccini and Huerta 2005). Although there are many successful applications of speech technology, there are domains where the results have not been satisfactory. One of these issues that concerns the current thesis is *naturalness* of synthesized speech, or of the overall human-machine interaction in general. This is discussed in more detail in section 2.2.3.

Beyond the current applications, speech technology is also an essential tool for an even greater goal: the understanding of *how human beings speak and understand speech,* and - more generally - *communicate.* The latter is the goal of speech and communication science, a diverse multidisciplinary field of research. Speech technology and speech science are connected bi-directionally: speech technology provides a test-bed for experimental testing of various existing and emerging speech science theories, while it borrows findings from speech science in order to achieve better performance in speech technology applications.

From the application point of view, speech interfaces are seen as potentially the most efficient possible, as speech is the most natural form of human communication (Lustgarten and Juang 2003). Thus, the development of naturally interacting speech interfaces points to better efficiency in current systems, as well as to an extension of the application field to tasks for which the current technologies are inadequate (Dybkjær and Dybkjær 2004). This is discussed further in the following

2 The first recorded "talking machine" (1769) was that of the Hungarian count Wolfgang Ritter von Kempelen, which was a mechanical apparatus that produced vocalic sounds. [http://en.wikipedia.org/wiki/Wolfgang_von_Kempelen%27s_Speaking_Machine,(01/04/2010)]

3 http://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_%28film%29,(01/04/2010)

section.

2.2.2 Recent trends in SDS research

One of the research areas in spoken language technology that has attracted interest in recent years is that of spoken dialogue systems (SDS). These have been successfully used in telephony, where they are also termed *interactive voice response* (IVR) systems. Automated call-routing (Williams and Witt 2004) is perhaps the most widespread use of this technology. These systems have been traditionally perceived as machines that understand spoken commands and produce monotonous spoken output. However, advances in natural language processing and increased computational capabilities have fuelled more optimistic visions and goals, moving away from the view of the computer speech interface as a tool, towards speculated "intelligent dialogue systems" and "communicative agents" (Jokinen 2000). The field of applications that such systems are thought to accomplish in the future is virtually endless (Jokinen 2003):

> "Sjöberg and Backlund (2000) envisage the future information and communication systems contain computers that are built into products such as clothes, books, beds, and sporting gear, and which communicate easily with other objects. Computers will also have senses and they can interpret human expressions, can smell, feel, hear, see and taste, and there will be intuitive human-computer interfaces that mimic human communication."

Indeed, talking agents are nowadays perceived both as machines and as "virtual persons" giving rise to the distinction between the *human metaphor* and the *interface metaphor* (Edlund *et al.* 2006), also explained in (Carlson *et al.* 2006):

> "In the interface metaphor, the spoken dialogue system is perceived as a machine interface – often but not always a computer interface. Speech is used to accomplish what would have otherwise been accomplished by some other means of input, such as a keyboard or a mouse. In the human metaphor, on the other hand, the computer is perceived as a creature (or even a person) with human-like conversational abilities, and speech is not a substitute or one of many alternatives, but rather the primary means of communicating with this creature."

This concept of human-like behavior extends to many applications, such as games and educational programs (Hakulinen and Turunen 1999). Moreover, research in emotional speech has led to the ambition of developing systems that can recognize and express emotions (Austermann *et al.* 2005; Lee and Narayanan 2005). As pointed out in (Holzapfel *et al.* 2002):

> "For example, uses have been software to assist learning and intelligent agents. It proved to

be beneficial for tutoring agents and learning software to show emotional behavior (e.g. the persona-effect) and use strategies based on emotional intelligence. For example motivating the user depending on his current emotional state [...]. Emotional intelligence has also been used in programs to improve user acceptance. This can be achieved by responding to user frustration and trying to help relieve frustration and recover to a positive emotional state […]. However, most applications are entirely unaware of the emotional state of the user and have no user model at all. This prevents a variety of possibilities to create programs that are better adapted to the user than today's programs are."

But, is this human-like "creature" the ultimate goal in speech technology research? Indeed, there are considerations against such an idea (Edlund *et al.* 2008), never mind whether it is even possible to ever build a "Turing machine"[4] (Larsson 2005). Why would we need a machine that is designed to perform tasks to be - or at least to behave - like us? The answer is that human-likeness is likely to enhance certain applications. For example, commercial SDS are likely to be more pleasant if the user had the impression that they were actually speaking to a *person*, even if they *knew* that they were speaking to machine. Some argument towards this issue is given in (Carlson *et al.* 2006):

"We are aware that more 'natural' or human-like behaviour does not automatically make a spoken dialogue system 'better' (i.e. more efficient or more well-liked by its users). Indeed, we are quite convinced that the advantage (or disadvantage) of human-like behaviour will be highly dependent on the application. However, a dialogue system that is coherent with a human metaphor may profit from a number of characteristics of speech that are typically not exploited in current systems designed with the interface metaphor in mind: it comes natural to us; it is good for reasoning and problem solving; and it is commonly used for social and bonding purposes, to mention a few."

(Edlund *et al.* 2009) points out that some of the benefits of using speech as an interface (works in hands-free and eyes-free situations or when other interfaces are inconvenient; provides an alternative interface for the disabled; or uses simple hardware such as a telephone) are more consistent with the interface metaphor and have been exploited accordingly in suitable domains (call routing, travel booking, voice commands, dictation, TTS text-reading). Other benefits, however, (simplicity, as humans are used to communicate through speech; quickness, as speech is fast to produce and perceive; robustness, utilizing human-like error-handling; and pleasantness, as human dialogue is sociable and pleasant) are more consistent with the human metaphor and have hardly been considered outside research systems (games and entertainment; education and learning;

4 Alan Turing (1912-1954) proposed a test, in which A (a human) has a dialogue with B (a machine). If A can be convinced that B is human, then B should be considered as having intelligence equal to a human (Larsson 2005)

expert systems for problem solving tasks such as IT support; and guiding, such as city-guides).

It should be made clear from the above that there is a definitive turn in spoken dialogue interface research, towards a mode of interaction that resembles everyday conversation between humans, i.e. towards *natural* interaction. However, this has proved to be a non-trivial undertaking:

> "The computer synthesis of natural-sounding speech has been a goal of computer scientists and speech technologists for more than half a century [...] yet neither linguists nor phoneticians have yet achieved a comprehensive definition of the full range and variation of speech as a means of human communication and social interaction.", (Campbell 2006)

Indeed, human speech and communication is characterized by complex phenomena that speech science is striving to explain. As a result, speech technology interfaces are characterized by lack of naturalness, or dissimilarity in comparison to the type of interaction that humans are accustomed to in their everyday life. The next section discusses the issue of naturalness in speech technology.

2.2.3 Naturalness in speech technology

As pointed out in the previous section, the issue of naturalness is not new in speech technology and has been one of the major goals (Campbell 2006) as well as one of the major short-comings in speech technology research. This section discusses the issue of naturalness in relation to various areas of speech technology.

The definition of naturalness in the field of speech technology has always been left vague. This is because providing a definition for naturalness raises philosophical questions, due to its subjective nature. Spoken language is ever changing in form and what appears as natural to one may appear unnatural to another. Different gender, age, cultural or ethnic groups use language differently. From a speech interface point of view, the properties of the system need to match the *expectations* of the user (Perez-Quinones and Sibert 1996) as to what is natural or not; given the diversity in expectations among the possible users, it becomes clear that this is a major problem. The typical answer to this problem is the adoption of a vague working definition of naturalness as a "convincingly human-like property" or "human-likeness" (Edlund *et al.* 2008).

The problem of evaluating how natural the synthesized speech sounds has not been solved either. The most frequently occurring solution is that of listening tests in which subjects are asked to rate the naturalness of the output, using scores on a scale from 1-10, on a number of different questions, a procedure that constitutes a *mean opinion score (*MOS) test (Dutoit 1997; Tatham and Morton 2005). An alternative to subjective testing is that of comparing the human subjects' *response* to

synthesized speech, against their response to natural (human) speech (Edlund *et al.* 2008). If the responses are the same, then the subjects can be assumed to perceive the synthesized speech as natural speech. This type of "objective" testing, which is possible only in a dialogue context, requires special care to ensure that the comparison of the responses is valid. This "equivalence" is not straightforward to accomplish, as discussed in section 2.2.4.

There are at least four major areas in speech technology where naturalness is a major issue. The first is text-to-speech synthesis (TTS). In the recent years, TTS has overcome the "intelligibility threshold" and efforts have been directed towards improving the naturalness of the speech produced by the various synthesis methods (Tatham and Morton 2005). In (Dutoit 1997) intelligibility and naturalness are presented as two "benchmarks" for synthesized speech. (Tatham and Morton 2005) points out that the two are not uncorrelated, as it is natural for human speech to be intelligible under realistic conditions (outside the laboratory), while synthesized speech is often unintelligible in such conditions. Prosodic modeling is perhaps the most significant improvement on the naturalness of synthesized speech, as it accounts for appropriate tone configuration of an utterance and alleviates the "robotic" sound of synthesized speech (Dutoit 1997).

The second area is *Emotional Speech Synthesis* (Schroeder 2001), which involves synthesizing speech that is expressive and conveys human emotions. Within the area of emotional speech synthesis, naturalness refers to the final output speech, or whether the intended emotion is conveyed in a natural way, so that it can be perceived as such. Emotional speech synthesis offers a significant improvement on naturalness of synthesized speech, as human speech conveys emotional content which is an essential part of human interaction.

The third area is that of recorded speech corpora. This area relates to the previous two as part of the development process, or as a "live" component of the system. In TTS, a corpus is required in order to synthesize new utterances, at least in the most successful concatenative and unit selection methods (Dutoit 1997): the quality of the corpus directly affects the naturalness of the TTS output, by providing coverage for all possible utterances that the system is designed to generate[5]. In emotional speech synthesis, corpora are required in order to obtain *acoustic correlates* of human emotions, or properties of the speech signal that are associated with a particular emotional state (Murray and Arnott 1993) or emotion "dimension" (Schroeder 2001). In the former case, it is required that the corpora contain a range of emotional states, which are appropriately labeled by expert listeners. In the latter case, the emphasis is placed upon spontaneous speech, which conveys genuine emotions and attitudes which are representative of real-life conversations. In both cases,

5 It is outside the scope of this thesis to describe TTS methods. For a review see (Dutoit 1997)

naturalness refers to the recorded (human) speech itself and – in particular – to the genuineness of the emotions with respect to their similarity to real life scenarios (Batliner *et al.* 2000).

The fourth and final area is that of spoken dialogue systems (SDS), where naturalness usually refers to the overall interaction, although it is not uncommon for it to imply only a part or component of the system as being "natural" (dialogue management, lexical choice, response time, voice tone, voice expressiveness etc). As SDS encompasses other areas of speech technology in its components, the overall naturalness of an SDS depends on the naturalness of its individual components (e.g. of the TTS voice). An important issue is that several of these component technologies have been developed with monologue speech in mind and are thus inadequate in a dialogue context (e.g. prosodic models – see section 2.4.4). The inadequacy arises from the fact that dialogue speech has properties not exhibited in monologue speech.

One such property is inter-speaker accommodation. Therefore, inclusion of this property in spoken dialogue interfaces is a possible path of improving the naturalness of such systems. Since human dialogues are characterized by complexity, due to its numerous properties, it is arguably useful to build upon current systems incrementally, by identifying a property in human speech, and evaluating its improvement on naturalness in human-computer interaction. An existing methodology for performing this task (Edlund *et al.* 2008) is described in the following section.

2.2.4 Evaluation of naturalness in SDS

(Edlund *et al.* 2008) proposed a procedure for implementing and evaluating individual properties of human interaction in SDS. This section outlines the key points of this procedure, which will be from here on referred to as the *human metaphor paradigm.*

Human users can be trained to use a system, such as an SDS, by learning its instructions one by one, but it is easier for them to understand the operation of a system (in general) through a *metaphor.* (Edlund *et al.* 2008) extended this idea to SDS, in that the design of a system can help the users perceive the system through a specific metaphor. Two contrasting metaphors were presented in (Edlund *et al.* 2008): the human metaphor and the interface metaphor (see section 2.2.2 for a description). Some users can better understand the operation of a system through the interface metaphor, while others can use a system more efficiently if they perceive it as having human-like abilities in speech production and understanding. In addition, the task a system is designed for can largely influence the type of metaphor that is more suitable (see section 2.2.2), while for some tasks both metaphors can be used. One can even imagine other metaphors that are in-between the two extremes, such as the "android" metaphor proposed in (Edlund *et al.* 2008), which are human-like

in some aspects and "machine-like" in some other aspects. (Larsson 2005) described a continuum defined between the "engineering" (interface metaphor) and "simulation" (human metaphor) positions.

In order to make human-machine interaction more *human-like*, which (Edlund *et al.* 2008) adopted as a working definition of "natural", an evaluation target is required. In this case, the evaluation target is human dialogue. The implementation evaluation schema is shown in Figure 2.1. The left-hand side of the picture depicts a human dialogue between two persons, h_1 and h_2. After measuring some property in the speech of h_1 and h_2, the property is implemented in C_1 on the right-hand side of the picture, which depicts a human-computer interaction. The implementation can then be evaluated in terms of (a) *similarity,* between the behaviour of C_1 and that of h_1 and h_2, and (b) *response,* if the behaviour of H_1 ,who interacts with the system C_1, resembles that of h_1 and h_2.

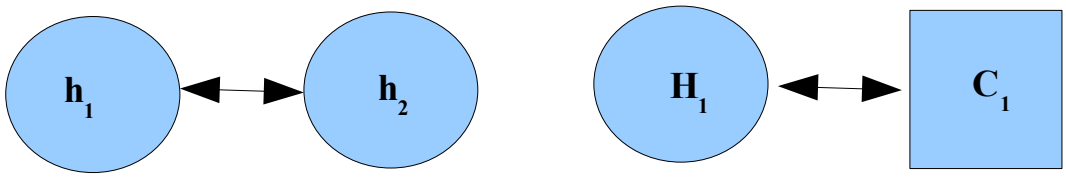


Figure 2.1: Schematic of human-human and human-machine interaction, adapted from (Edlund et al 2009)

In this manner, the implementation can be evaluated both subjectively and objectively: (a) above can be evaluated perceptually in listening experiments, in which subjects judge whether the behaviour is firstly perceivable, and secondly whether the behaviour of the system resembles that of a human more than a control condition in which the behaviour is not present, but (b) can be directly evaluated by measuring the properties of the user speech (in the case of human-computer interaction), and comparing the results to those from the human-human interaction. As pointed out in (Edlund *et al.* 2008), this comparison is not always straightforward, and the equivalence must be considered carefully, by keeping as many variables as possible constant in the two cases. Three key methods are presented for extracting speech features in a way that can enable such comparisons: micro-domains, direct data manipulation of human dialogues and Wizard-of-Oz experiments.

Micro-domains are interactions that are constrained in such a way that human-machine interaction *can* be perceived as human-like. (Edlund *et al.* 2008) provided the example of narration, a type of interaction in which the listener is not expected to interrupt. A state-of-the-art unit-selection speech synthesis system in such a limited domain is very likely to be perceived as an actual human speaker. According to (Edlund *et al.* 2008), the usefulness of micro-domains is that they can be used to

directly model the user behaviour in some respect. In the case of narration, this can be backchanneling feedback responses which signal attention.

Direct data manipulation refers to changing the speech properties of one of the speakers in the human dialogue and using the resulting modified signal as the "system" voice. A distinction can be made between on-line and off-line manipulation. In off-line manipulation, the speech signal is altered is some way (such as prosodic modification or introduction of longer silence before utterances) and presented to subjects in perceptual listening tests, in order to accumulate judgements on the effect of the manipulations. The drawback of this method is that it cannot be used to measure the user's *response* to the manipulations, because introducing changes to the speech of one speaker arguably changes the interaction in such a way that the data from the second speaker is no longer valid: had the manipulation occurred *during* the interaction, the speech of the second speaker would have been different. This problem is not present in on-line manipulation. The latter method also has the advantage of using data from both speakers, if the manipulations are made symmetrical. However, the level of control is lower than that of Wizard-of-Oz experiments (see below), because computationally expensive manipulations introduce latencies to the interaction. (Edlund *et al.* 2008) provided only one example of this method actually being used to manipulate speech (noise contamination of the signal in order to elicit acknowledgement requests), while it has also been used to manipulate text-chat and gestural features.

Wizard-of-Oz experiments (Woffit *et al.* 1997) are simulations of functioning systems, in which subjects are led to believe that they are interacting with a fully automated SDS, while – in reality – a human experimenter is controlling some aspect of the system. These have been used for several purposes, such as in the research and design phase of many SDS (Edlund *et al.* 2008). Experiments of this type also present a viable option for evaluation of human-likeness in SDS. Since unlimited conversational SDS are currently unavailable, a Wizard-of-Oz set-up can be used instead, in order to monitor the user perception and response to an implemented human-like property in the interaction. This is possible both by use of questionnaires (or any other method of recording user judgement), as well as by directly measuring properties of the users' speech. Another advantage of this method is that it allows a significant level of control, in terms of manipulating the interaction in order to elicit a particular response from the user.

The proposed evaluation framework of (Edlund *et al.* 2008) makes it possible to implement and evaluate models of various aspects of natural human speech. Inter-speaker accommodation is one of these aspects that (Edlund *et al.* 2008) identified as a possible target for such an implementation. This is desirable for enhancing the "human metaphor" and improving on the human-likeness or

naturalness of the interaction in general. According to the outlined methodology of (Edlund *et al.* 2008), this requires careful investigation and characterization of accommodation phenomena in human speech, in order to inform SDS design strategies that can take them into account.

Currently, there are two major obstacles to such an implementation. First, there is insufficient knowledge on accommodation phenomena, especially in relation to their *form*. As was mentioned in the introduction, traditional descriptions of inter-speaker accommodation lack a quantitative approach that can inform SDS implementations (Oviatt *et al.* 2004). Second, state-of-the-art SDS are not yet capable of human-like communication, nor are they likely to be in the near future (Larsson 2005). There are, however, components in current architectures that can benefit even from "naive" implementations of accommodation phenomena. These components relate to several aspects of dialogue (dialogue management, turn-taking, prosody, emotional speech) that have been identified among the major issues in recent SDS research discussions (Minker *et al.* 2006). The following section provides a brief outline of the operation and major components of SDS systems, discussing the implications of inter-speaker accommodation where appropriate.

2.3 Spoken dialogue systems

This section provides a brief description of spoken dialogue systems, in order to highlight areas in which inter-speaker accommodation may improve current performance. It is noted that this is not a strictly technical description, but rather a conceptual (abstract) discussion, from which several insights can be drawn in relation to possible improvements in naturalness of such systems. A brief outline of the operation of SDS is given in section 2.3.1. Section 2.3.2 discusses the issue of the floor-exchanging strategy of SDS, which is termed *interaction management,* in comparison to floor-exchanging in human dialogues as described by studies in conversation analysis. The conversational capabilities of SDS are discussed in section 2.3.3, under the topic of *dialogue management,* which is the central component of the SDS architecture. Finally, the issue of multimodality, which is the transmission of information through various parallel communication channels (lexical, prosodic, gestural) is discussed in section 2.3.4.

2.3.1 Operation of SDS

Spoken dialogue systems combine a number of other speech technologies, such as text-to-speech synthesis (TTS), *automatic speech recognition* (ASR), *automatic language understanding* (ALU), *voice activity detection* (VAD) and *natural language processing* (NLP). A schematic of the operation of SDS is shown in Figure 2.2.

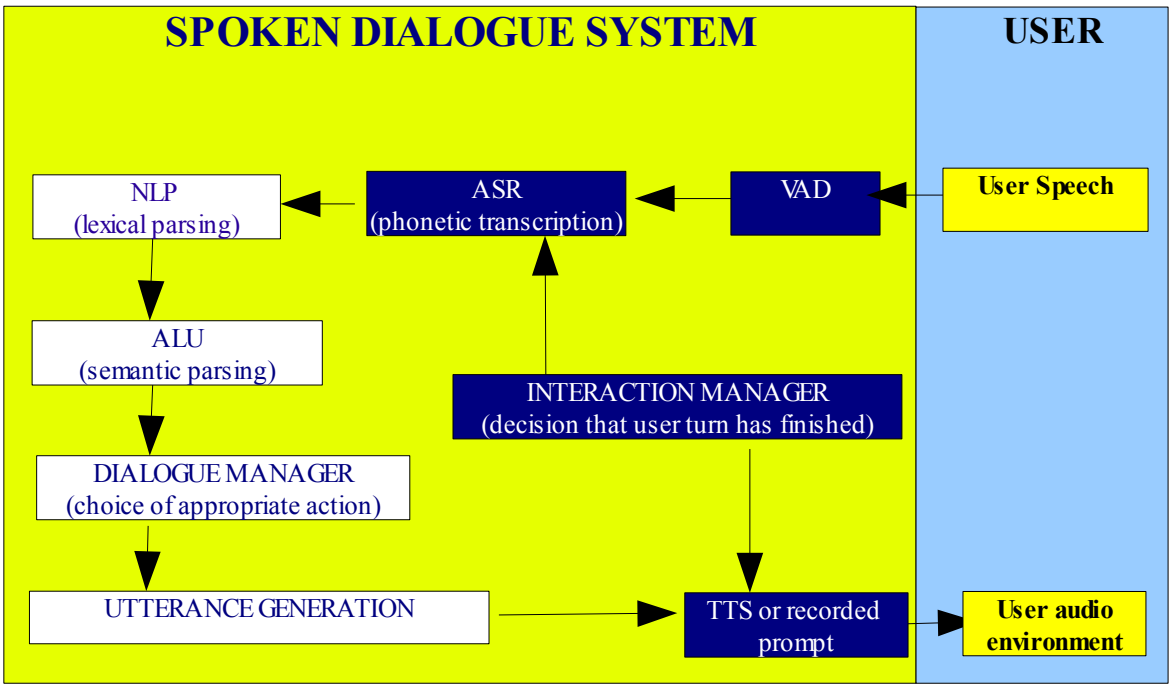


Figure 2.2: Schematic of spoken dialogue system

The interaction between a user and an SDS requires audio sensory equipment (microphone and speakers, or a telephone), unless visual information is also available, in which case visual equipment (camera, monitor) is also required. User utterances are detected by means of a *voice activation detection* (VAD) algorithm (Song *et al.* 2009). Spoken input from the user is processed when the user's utterance is completed. This decision is made by the interaction management component, typically by means of silence duration threshold (see next section).

If the user turn has ended, the recorded user utterance is passed to the ASR component, which transforms the recorded speech signal into a *phonetic transcription,* i.e. a series of phonemes. This transcription is input to the NLP component, which performs lexical parsing, identifying the lexical elements from a list of possible candidates and outputs a word stream. The latter is semantically parsed by the ALU component, so that the system can "understand" the utterance. The dialog manager decides on the appropriate action (provide a reply, ask for clarification etc.) and an appropriate utterance is generated (or chosen from a preset list of available prompts) and passed to the speech synthesizer in order to be "spoken" to the user.

The above is only a basic description of the operation of an SDS, but it is sufficient for the purpose of discussing some of the key issues in relation to inter-speaker accommodation. In particular, two areas are identified as currently limiting the human-likeness of human-machine interaction: the dialogue manager component, which is central to the SDS architecture and represents all the possible actions the system can perform (Pieraccini and Huerta 2005), and the interaction manager

component, which is responsible for smooth floor-exchanges between the system and the user (Raux 2008). In addition, the inclusion of additional communication channels, or modalities, has been identified as a key area of improvement on users' perception of the human metaphor (Edlund *et al.* 2008). The following three sections outline a number of key issues which are related to these areas in SDS research.

2.3.2 Interaction management

Traditionally, SDS have adopted a "push-to-talk" or "ping-pong" turn-talking strategy, in which there is a rigid one-speaker-per-turn succession between user and system (Carlson *et al.* 2006). All this requires is the detection of the end of the user turn, also termed *end-pointing detection*, or simply *end-pointing*, a process based on a silence duration threshold, typically from 500-2000 ms (Edlund *et al.* 2005). This approach introduces false alarms, when the user hesitates, or unwanted latencies, when there is an actual endpoint and the system "waits" for a time equal to the duration threshold. This problem arises from the "ping-pong" view of dialogue, which is not consistent with actual everyday dialogue between humans (Furui *et al.* 2005). Thus, research has recently turned towards conversation and discourse analysis (Mushin *et al.* 2003) in order to implement more adequate interaction management strategies in spoken dialogue systems.

A famous quote from the seminal paper on turn-taking (Sacks *et al.* 1974) states that "in any conversation, we observe the following: speaker-change recurs, or at least occurs. [...]". Sacks et al proposed perhaps the first systematic account of how turns are exchanged, inaugurating the field of conversation analysis (Raux 2008), which evolved into discourse analysis (Campbell 2009), although much earlier *chronographic* records of dialogues are reported in (Lennes and Anttila 2002), and (Campbell 2009). In the model of (Sacks *et al.* 1974), turns are defined by means of turn-construction units and turn-allocation units. A central concept is that of transition-relevant points (TRP). A TRP is a point in the dialogue where potentially there can be a turn-exchange. (Raux 2008) presented previous research on several TRP cues, which include syntactic conclusion, prosody, semantics/pragmatics and non-verbal behaviour, such as making eye contact in order to indicate the end of a turn. According to Raux, the only objective definition of a turn that does not take into account interpretations by the researcher (which would lead to subjectivity) is that of (Jaffe and Feldstein 1970):

> "The speaker who utters the first unilateral sound both initiates the conversation and gains possession of the floor. Having gained possession, a speaker maintains it until the first unilateral sound by another speaker, at which time the latter gains possession of the floor."

However, natural human dialogue does not consist of a mere "exchange of turns", but there are many instances of overlapping speech, which serve as acknowledgments of continuing attention, agreement or may be attempts to interrupt and take over the floor. A definition of overlapping speech segments, or *overlaps,* is given in (Delmonte 2005):

> "Overlaps may be defined as a speech event in which two people speak simulateneously by uttering actual words or in some cases non-words, when one of the speakers, usually the interlocutor, interrupts or backchannels the current speaker."

Among the first to view conversation as a "collaboration" were Clark and Schaefer (1989), who defined the conversation as a joint process between two partners who join a "pact" with some prior knowledge, obligations, and goals. The way of achieving these goals is by means of *contributions,* in order to establish *common ground* (shared knowledge). A central concept in discourse analysis is that of *speech acts* or *dialog acts* (Wright 1999), which are a categorization of all utterances in a dialogue, with each act serving a distinct communication purpose in the discourse. For example, short, overlapping utterances which can be anything from hums and noises to utterances such as "yes, yes", or even longer utterances, are used in spoken dialogs to signal acceptance or disagreement, or can be prompting the speaker for continuation/interruption of their current turn. *Back-channel feedback* or *back-channeling* is a term coined for these dialog acts that serve the double purpose of conveying the listener's attitude towards what the speaker is currently saying and managing the transition of turns. However, there is still no consensus on what a turn is, never mind a categorical description that can be used to annotate speech corpora in a straightforward way (Bosch *et al.* 2005; Raux 2008).

Further, the view of conversation partners having distinct roles of "speaker" and "listener" are fictional according to some (cf Heylen 2009): interlocutors do not take turns to speak, but rather accommodate the transition from one speaker to the next by means of cues. More recently, there have been proposed representations of human dialogue as a joint process, in which both participants are actively participating *continuously* through the process of *active listening* and *synchrony* (Campbell 2009), *feedback* and *instantaneous response* (Heylen 2009). The former study is closer to the accommodation phenomena line of research, as it focuses on synchronous behaviour of interlocutors, while the latter is based on a discourse analysis point of view. (Heylen 2009) suggested a schematic representation of dialogue (see Figure 2.3):

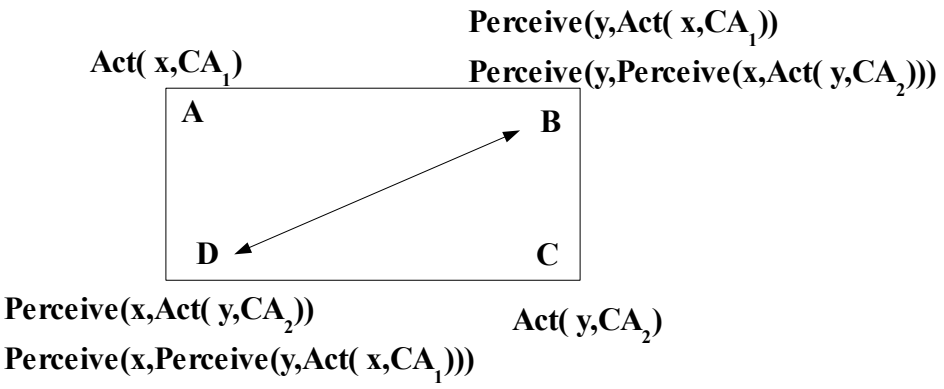


Figure 2.3: Schema of dialogue interaction adapted from (Heylen 2009)

In its simplest form, shown in the figure above, the schema essentially considers a *feedback loop* (A→B→C→D→A) together with simultaneous transmission of information from both speakers (diagonal line). A communicative act A by speaker x is perceived by speaker y, who produces his/her own communicative act as a response. This is perceived again by x and so on in this circular fashion. Thus the communicative acts of both x and y are influenced by the previous responsive acts of each other, resulting in a feedback loop. In addition to this, both x and y provide acknowledgements to each other that they have perceived the communicative acts. This is performed instantaneously by use of overlapping speech (back-channeling) or head nods, facial gestures, eye movements etc., in the case of face-to-face conversation. In this manner, x perceives the acknowledgement of y while x *is still producing their own utterance*, which again influences the communicative act manifestation. This also explains lexical and syntactic alignment, as well as utterance length accommodation (Matessa 2001) which are discussed in chapter 3. Thus, the schema of (Heylen 2009) captures both feedback and monitoring, two processes that accommodate interaction management in natural human dialogues.

Commercial systems systems, however, have largely remained loyal to silence-threshold turn-taking strategies, driven by such goals as task-completion efficiency and robustness (see section 2.3.3). For example, push-to-talk interfaces (that require the user to push a button in order to speak to the system) can be equally or even more efficient than free turn allocation interfaces for certain tasks (Fernandez *et al.* 2006). However, there are also significant developments towards more realistic turn-taking behavior in SDS, which take into account temporal, prosodic, syntactic, semantic and pragmatic (domain knowledge) information in order to improve end-of-turn detection while allowing for non-interrupting overlaps (e.g. Raux 2008).

Knowledge of inter-speaker accommodation phenomena could improve the performance of interaction management in SDS. For example, accommodation of pause duration between speaker

turns has been reported in several studies (Jaffe *et al.* 2001; Bosch *et al.* 2005; Edlund *et al.* 2009). Therefore, turn-taking behaviour could be improved by dynamically adjusting the silence threshold according to the on-going activity in the interaction, e.g. by monitoring silence duration before speaker turns and its variation according to silence duration before system turns.

2.3.3 Dialogue management

The gap that divides current spoken dialogue systems from human-like conversational speech is perhaps most evidently illustrated by the type of dialogue that current SDS are capable of entertaining. Currently, SDS are incapable of engaging in human-like conversation that exhibits spontaneous speech, however they *are* capable of dealing with increasingly complex tasks in commercial applications (Allen *et al.* 2001; Dybkjær and Dybkjær 2004). This is made possible by the *dialogue manager,* the SDS component responsible for controlling the interaction with a user. A categorization of SDS that is relevant to the organization of the dialogue is that of *initiative*: typically, applications such as call-routing or travel booking are *system-initatitive* (the system asks questions and idly waits for user replies). *User-initiative* refers to the opposite schema of operation, e.g. a user articulates queries to a database. *Mixed-initiative* systems (Allen *et al.* 1999) can combine both approaches, either in presenting an open prompt in order to circumvent the requirement of presenting all possible menu options, or as an adaptive strategy, depending on the dialogue flow (Litman and Pan 2002).

According to a review of existing dialogue management implementations (Pieraccini and Huerta 2005), commercial systems and research on SDS have followed contrastingly different routes: spoken dialogue interaction research aimed for "conversational interfaces", and fell back to more feasible goals when limitations became apparent. In contrast, commercial systems followed a "bottom-up" evolutionary path, as a result of designing SDS for specific applications, in which the domain constraints made a speech interface feasible with the technology that was available. In flight booking, for example, the dialogue – even with a human agent - follows a strict procedure in which all fields in a form have to be filled before a booking can be completed.

The beginnings of dialogue management in the commercial domain were *directed dialogue* systems, in which a sequence of prompts was presented to the user and resolution of each step was required for the script to proceed to the next action. These were typically developed directly on the application platform and used proprietary development tools, resulting in zero portability and re-usability. The next step was finite-state machine dialogue modeling (Pieraccini and Huerta 2005). In this paradigm, a dialogue is represented by a flow diagram of nodes and arcs. The nodes represent states in the dialogue (such as waiting for a specific user input) and the arcs represent transitions

between states, depending on conditions. The simplest form of this is a call flow diagram (such as the flight booking example above). The advantage of this method is that finite state machines can be re-used for several applications, while the *topologies*, or task-specific requirements can be accommodated by adapting the flow diagram. A limitation of this method is that all possible outcomes in each dialogue state must be thought of in advance, which makes management of more complex tasks (such as problem solving) impossible to program.

A further improvement on finite state dialogue managers was the abstraction of states and arcs into a functional control manager, which is a finite state representation of the dialogue control logic (rather than the dialogue itself). In this case, states correspond to functions that are executed depending on conditions that are evaluated by means of separate memory structures, which can be accessed by all states. In this manner, more complex topologies can be modelled. However, complex tasks (such as problem solving) can result in very complex models which are impractical or insufficient (Dybkjær and Dybkjær 2004). This problem has not yet been resolved in the commercial domain, but there are approaches in the research community towards resolving these problems (Allen *et al.* 2001). Inference based dialogue managers make use of domain knowledge and strategy, which is defined as goals and sub-goals. These can be re-defined dynamically during the dialogue, thus giving rise to the term "adaptive" dialogue management. The advantage of this approach is that it provides the dialogue manager with a set of actions it can perform, without having to define every state and transition separately, thus allowing for more complex topologies to be implemented (Pieraccini and Huerta 2005).

Despite the advances in dialogue management described in (Pieraccini and Huerta 2005), SDS are still incapable of conversing in a "human-like" manner (Dybkjær and Dybkjær 2004) and there are arguments against the idea that this goal will ever be feasible, as it would require computer agents with human *intelligence* (Larsson 2005). However, (Larsson 2005) points out that this does not mean that spoken dialogue research is without purpose, as it can still largely improve SDS in terms of naturalness.

One of the most important issues in dialogue management is error detection and an error-recovery startegy. i.e a "fall-back" plan when things go wrong in the conversation (Carlson *et al.* 2006). Error detection is crucial for the operation of SDS, as failure to recognize an error can result in either acceptance of erroneous input, or user frustration (or both). Error detection strategies typically utilize confidence scores (Lee and Narayanan 2005) from the ASR component (a low score is indicative of a possible error). The semantic parser may also provide error detection functionality if it fails to understand the user request. It is also possible to detect errors with the help of prosody, as

user rephrased or repeated commands have been found to have different prosodic content from utterances in smooth regions of the interaction (Bell *et al.* 2003), or by means of emotion recognition (see section 2.4.4). In addition to detecting errors, an SDS must have a *recovery* strategy. In the simplest of implementations, this can be a clarification request in the form of yes/no question (which the system can recognize with more confidence), or, in case of severe errors, dispatching the task to a human operator (Pieraccini and Huerta 2005).

A quantitative description of inter-speaker accommodation can enhance the user monitoring and error detection capabilities of SDS, as they are based on an on-line analysis of features extracted from the speech signal. For example, a/p features can be used to determine possible user frustration, based on online monitoring of user emotions (Holzapfel *et al.* 2002), as emotion recognition is based on classification based on these features. This classification can perhaps be improved if variation of prosodic features due to inter-speaker accommodation is taken into account.

2.3.4 Multimodality

In human interaction, information is exchanged through various modalities: lexical content, syntax, prosody, facial expression, gesture, gaze and "body language". It is noted that "information" in the context of interaction denotes either pragmatic content, or expression of one's emotion, attitude or belief on a particular subject. These communicative functions are expressed *simultaneously* through the various modalities. For example, pleasure/displeasure on a particular situation that is being discussed may be expressed lexically, but this is often accompanied by manifestations of this mood in other modalities (e.g. smile or disgusted facial gesture, relevant intonation, possible hand gestures). This *multimodality* is an intuitively known property of human interaction.

Spoken dialogue systems have utilized multimodality (Dybkjær *et al.* 2004; Oviatt *et al.* 2004; Pieraccini *et al.* 2009) through the inclusion of *avatars,* which typically have the form of animated talking heads (McTear 2004). These exhibit impressive capabilities in terms of lip-synchronization with the speech signal, as well as displaying facial gestures and nods of approval or acknowledgment (signaling understanding). The inclusion of talking heads is considered as a significant improvement on the naturalness of the interaction with SDS and has been used successfully in various applications (Pieraccini *et al.* 2009). Other modalities are the use of light pen or hand-writing, 2D gesture input and graphics – such as images and maps – output (Dybkjær *et al.* 2004).

As mentioned in the introduction, and further discussed in chapter 3, inter-speaker accommodation is known to occur along different modalities simultaneously. Therefore, an implementation of

accommodating behaviour in different modalities in an SDS environment is likely to bring improvements to both naturalness and efficiency. This has been proposed in (Bell *et al.* 2000), which involved subjects interacting with a multimodal SDS. (Campbell 2009) demonstrated simultaneous activity of interlocutors across several modalities (speech, head movement, body movement) in spontaneous (human) dialogues. Implementation of similar behaviour in SDS would intuitively enhance the perception of naturalness.

2.4 Prosody

As was mentioned in section 2.2.3, the study of speech prosody has resulted in major improvements of naturalness in speech technology. In addition, sections 2.3.2 and 2.3.3 have already identified areas in SDS research in which prosody plays an important role. This section presents a sufficiently detailed account of the form (section 2.4.1) and function (sections 2.4.2 - 2.4.3) of prosody, in order to further illustrate its importance in dialogue systems and the motivation for studying accommodation of prosodic features.

The word "prosody" is of ancient Greek origin: According to Diomedes (400 BC), prosody "is sung with the syllables", an etymological definition referring to the strict rhythmic and melodic rules (similar to music) of ancient Greek, hence the Latin equivalent ac-centus (ad –cantus): accent. Therefore, prosody refers to the melodic (pitch) and temporal (speech rate, phoneme duration) features of speech. However, since these features are studied from different points of view, such as linguistic studies and engineering applications, there is no universal definition for prosody (Cutler *et al.* 1997). (Werner and Keller 1994), for example, examining prosody from a speech technology (synthesis and recognition) point of view, re-stated a classic definition of prosody as "the speech features whose domain is not a single phonetic segment, but larger units of more than one segment, possibly whole sentences or even longer utterances". (Werner and Keller 1994) thus accepted the equivalence of prosody to *suprasegmental* features, a term attributed to (Lehiste 1970). This equivalence is also present in the definition of (Dutoit 1997):

> "The term prosody refers to certain properties of the speech signal such as audible changes in pitch, loudness, and syllable length. […] are also referred to as *suprasegmental* phenomena"

According to both (Dutoit 1997) and (Werner and Keller 1994), there is a number of different representations or "levels" of prosody. (Dutoit 1997) distinguishes three different representations (see Table 2.1). The *acoustic* level refers to measurable properties of the speech signal, such as fundamental frequency (F0), amplitude and segmental duration, which is included despite not being a strictly acoustic feature.

Acoustic	Perceptual	Linguistic
Fundamental Frequency (F0)	**Pitch**	**Tone, intonation, aspect of stress**
Amplitude, Energy, Intensity	**Loudness**	**Aspect of stress**
Duration	**Length**	**Aspect of stress**
Amplitude dynamics	**Strength**	**Aspect of stress**

Table 2.1: Three representations of prosody and their properties (Dutoit 1997)

The *perceptual* level refers to perceptible features of prosody. Fundamental frequency is an acoustic property of the signal that is perceived as *pitch*. Similarly, intensity, amplitude or energy can be perceived as *loudness*. Thus, a variation at the acoustic level has to be large enough to be perceived as such at the perceptual level, and micro-perturbations of the same acoustic features should not be (mistakenly) characterized as prosodic variations. The properties of the perceptual level have, in turn, their own correspondences to the properties of the *linguistic* level. In particular, *intonation* can be associated with pitch (and therefore F0). On the other hand, acoustic correlates of *stress* have been difficult to define (Dutoit 1997).

The general consensus is that F0, intensity, rate of delivery and duration are the most important prosodic features (Hakulinen and Turunen 1999). Other signal features have been considered as prosodic, because they satisfy the definition above of being “suprasegmental” (spanning several segments). One such example is voice quality (Laver 1980), which is "the characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and supra-laryngeal features and running continuously through the individual's speech" (Trask 1996). However, the four features mentioned above are widely accepted (Dutoit 1997) as the most relevant in the study of prosody:

a) Pitch/F0: Human speech is a *quasi-periodic* signal (in voiced regions). The vibration of the vocal folds in the larynx, coupled with the resonances (or formants) of the oral and nasal cavities, produces the voiced speech sounds (vowels and voiced consonants), while passage of the turbulent breath stream through narrow constrictions in the oral cavity produces fricatives (e.g. */s/* and */f/*), and sudden releases of built-up air pressure, modulated by constriction produce stop consonants (*/p/*, */k/* and */t/*) (Pickett 1999). The vibration of the vocal folds exhibits micro-perturbations, i.e. consequtive periods have slightly different length. This phenomenon is termed jitter (Titze 1994). However, under normal circumstances, these

perturbations are small and any voiced sound can be approximated by a periodic signal, with a fundamental frequency (F0) measured in Hertz or semitones. The term F0 contour or pitch contour refers to a continuous curve that represents the variation of F0/pitch over a given amount of time (Dutoit 1997). Thus, a segmental F0 contour describes the shape of a pitch accent or tone, while a phrase or sentence pitch contour describes phrase or sentence intontation, respectively (see Figure 2.4).

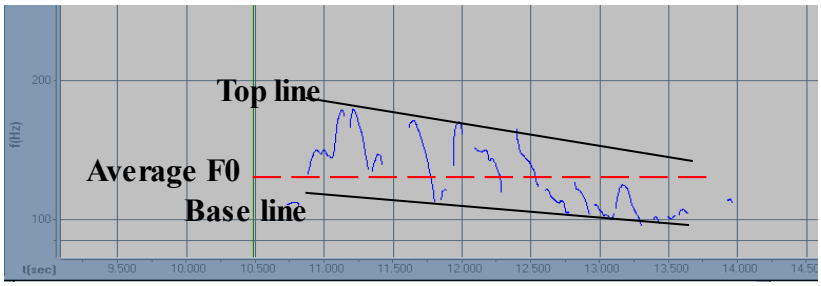


Figure 2.4: Utterance F0 (pitch) contour with stylization (topline, baseline) lines and average F0 line shown

b) Intensity: In signal processing terms, the intensity of a sound wave is the average amount of energy transmitted per unit time through a unit area in a specified direction (Pickett 1999). A simpler description would be that intensity expresses the amount of pressure or energy that a travelling wave carries. A speech sound with higher intensity is perceived as being "louder". Intensity is measured in *decibels* (dB) relatively to a reference pressure, thus yielding a *sound pressure level* (SPL). For speech sounds, the reference pressure is the *auditory threshold*[6] ($2 * 10^{-5}$ Pascal) and thus the value of intensity in dB SPL is given by the equation:

$$I = 20 \times \log_{10} \frac{P_{sound}}{P_{reference}}$$

Equation 2.1: Intensity of a speech sound in decibels (dB)

where I is the intensity, $P_{reference}$ is the auditory threshold, and P_{sound} is the sound pressure in Pascal. Intensity in speech exhibits micro-pertirbations, termed *shimmer* (Titze 1994), which are again negligible in the study of prosody. An intensity contour is a continuous line that represents the variation of the speech signal intensity over time (Dutoit 1997).

c) Speech rate. The rate of delivery, or speed of delivery, or speech rate of the signal expresses

6 20 μPa in air and 1 μPa in water (ANSI S1.1 – 1994) is the minimum sound pressure noticeable by a young person with undamaged hearing, for a pure sine wave tone with a frequency of 1000 Hz.

how fast an utterance is spoken. This measurement can only be an approximation, since it is usually calculated over the length of an utterance, under the assumption that it remains constant during that period. Therefore, it is actually a *mean velocity* that is usually calculated in syllables per minute or vowels per minute (Pfau and Ruske 1998; Wang and Narayanan 2005). The vowel length is also (negatively) correlated with speech rate, as faster delivery implies shortened vowels. However, this method of calculation is unreliable, since vowel length is also subject to variations not related to speech rate (Galanis *et al.* 1996).

d) Segmental duration. The length of a speech segment is measured in milliseconds (ms). *Phonetic segmentation* is a process of identifying phoneme boundaries in a recorded speech signal. This can be done manually, which is a tedious process, or automatically (e.g. in ASR), using an algorithm. Both approaches are subject to a certain amount of error, although results from automatic segmentations have become reliable (e.g. Chang *et al.* 2000) .The length of syllables or vowels is typically related to prosodic effects: vowel lengthening is an acoustic correlate of linguistic stress, but can also serve other functions (e.g. lengthening of the final vowel in an utterance signals continuation of the current speaker turn in a dialogue).

2.4.1 Prosodic modeling

In TTS, *prosodic modeling* aims to inform speech synthesis by deriving the prosodic (melodic and temporal) structure of a sentece from the textual input and/or any other information that is available, such as a prosodic mark-up on words or syllables that need to be emphasized (Dutoit 1997). Various different prosodic models have been introduced in decades of research. As reported in (Kochanski 2006), the majority have focused on intonation (the fundamental frequency contour of an utterance). Intonation models have been categorized (e.g. Botinis *et al.* 2001), in a way that corresponds to the categorization of (Dutoit 1997) for prosodic representations (see previous section). Table 2.2 shows such a categorization and the most representative models of each category. Prosodic modeling has significantly improved the naturalness of TTS, as synthesized speech without any form of implementation for prosody sounds robotic and monotonous (Dutoit 1997). An exception to this rule are unit selection synthesizers that use large “chunks” of recorded speech corpora as units (e.g. Chu *et al.* 2001), as these long units already carry a prosodic structure.

Despite the advances that prosodic modeling has brought to TTS, certain limitations have been encountered. These arise mostly from the complexity of prosody itself (the mapping problem), as well as the fact that the most prominent prosodic features (pitch, intensity, speech rate, duration) have several functions in language (the function problem). In addition, and perhaps most importantly, it has been argued that proper prosodic rendering for text input requires *intelligence*, or

rather “world knowledge” (Dutoit 1997), which is how human speakers can read aloud any text and use prosody to put emphasis where appropriate. Thus, it is not surprising that the most natural-sounding speech comes from limited domain unit selection synthesizers (Narayanan and Alwan 2004), where the content words and phrases that need to be emphasized can be more accurately predicted. However, there are applications (such as SDS) in which additional semantic information is available (domain knowledge). This information, combined with the textual input, can provide for more appropriate prosody in the final synthesized utterance, a methodology that is also known as *concept-to-speech synthesis* (Dutoit 1997).

Phonological	**Perceptual & acoustic stylization**	**Acoustic-phonetic**
Pierrehumbert's intonational phonology (Pierrehumbert 1980) **Ladd's phonological intonation model** (Ladd 1983, 1996) **TOBI** (Silverman *et al.* 1992)	**IPO** (t'Hart *et al.* 1991) **TILT** (Taylor 2000) **INTSINT** (d'Alessandro and Mertens 1995)	**Fujisaki model** (Fujisaki 1992)

Table 2.2: Categorization of the most important intonation models in TTS

The *mapping problem*, as described in (Taylor 1992), refers to the three different representations of prosody mentioned above and how each representation is reflected in a particular prosodic model. All models have three basic components or layers: (a) a phonology, which is an abstract representation of prosodic boundaries and accents and may or may not be informed by linguistic phonology, (b) an intermediate layer, which is the core of the model and typically comprises an inventory of abstract prosodic units, and (c) an acoustic realization of the prosodic structure into actual numerical values which constitute the input to the speech synthesizer. Depending on the representation to which the model is closest (and thus categorized in Table 2.2), at least one of the mappings between two of the layers becomes problematic. Pierrehumbert's model (Pierrehumbert 1980), for example, has a straightforward mapping between the abstract and intermediate form, as it is a linguistic model, but actually computing values for prosodic features requires several assumptions and arbitrary choices, for example in defining a pitch baseline. (Taylor 1992) points out that this is a case of an “one-to-many” mapping: several realized utterances share an identical

intermediate prosodic representation. Fujisaki's model (Fujisaki 1992), a purely acoustic model, is much better at computing values from the intermediate representation, but it is very difficult to assign linguistic meaning to the core elements of the model (phrase and accent commands).

So far this discussion has covered problems associated with the *form* of prosody, but the situation is equally problematic in respect to its *function*. It has been argued (Kohler 2004) that the majority of prosodic models have overlooked the functional aspects of prosody (e.g. Aubergé 2002). A first approach towards function-based descriptions of intonation was made in (Kohler 1991), by integrating semantics/pragmatics and expressive functions of intonation in the Kiel Intonational Model (KIM). The development of KIM was established on the discovery of meaningful functional contrasts related to the position of F0 peaks in accented syllables (early vs medial vs late peaks) in German. A later function-based approach is the PENTA model (Xu 2005). However, these approaches cover only some of the multiple functions of prosody in human speech. The following section gives a brief overview of these functions.

2.4.2 Functions of prosody in human speech

According to (Cutler *et al.* 1997), prosody has at least four distinct contributions to language understanding: At the pre-lexical and lexical level, prosody aids the listener in identifying word boundaries and recognizing words, by use of strong-weak syllable contrasts and stressed syllables. Especially in the case of tonal languages, the stressed syllable is essential in resolving the ambiguity that arises from many possible words that only differ in their stress. At the structural level, prosody provides cues that aid the listener infer the syntactic structure of the utterance, although the mapping between the prosodic and syntactic structures is not *isomorhpic*. Finally, prosody is strongly related to understanding at the discourse/pragmatic level, where focal stress is used to distinguish newly introduced from already known information, or to resolve ambiguities and emphasize the important part of a sentence (e.g. "Mary did not come to Dublin by plane", where putting stress on each of the underlined words emphasizes a totally different point).

At the signal (acoustic) level, pitch (or F0), intensity, speech rate and duration are the speech features associated with the above *linguistic* functions. However these acoustic/prosodic (from here on a/p) features, carry several other *paralinguistic* functions (Kochanski 2006). Paralinguistic communication refers to aspects of speech that are not parts of the language or its spoken, verbal form, but are nonetheless required in order to communicate a speaker's affective state, attitude, or emotion, or to regulate time-sharing of the conversation.

(Gussenhoven 2005) distinguished three paralinguistic "codes". These are the frequency, effort, and

production code. Frequency (or pitch) is primarily associated with the size of the larynx and, therefore, with the speaker's age and gender. As an extension of this pitch-based biological distinction, the frequency (pitch) of speech can be used to signal masculinity or femininity, dominance or submission, friendliness or hostility, vulnerability or protectiveness. These distinctions are, according to (Gussenhoven 2005), related to the biological or cultural roles of genders and/or primal codes of behaviour of humans and animals. A lower pitched voice indicates a longer larynx, i.e. a larger animal, which can be more aggressive or dominant.

The effort and production codes are associated with the energy required to produce the speech signal. In particular, the effort is represented by variations in the pitch span, while the production code is associated with pitch and loudness declination due to the correlation of utterances and breath groups. Effort can be used to focus on significant parts of the utterance through various mechanisms such as wider pitch span or delayed peaks. Articulation precision that is higher than average in the utterance is another manifestation of the effort code. According to (Gussenhoven 2005), the effort code can be used to signal surprise or concern. The production code is better demonstrated in utterances that present variations in their normal declination trend. A higher than usual initial tone can indicate the start of a new topic, while a high or low final tone can respectively signal continuation or finality, allowing the listener to assess the information or respond before proceeding further (Gussenhoven 2005). Therefore, prosody enables dialogue organisation and smooth transitions between speakers engaging in conversation. This is discussed further in section 2.4.4.

2.4.3 Prosody and emotions

Modern research in speech analysis and synthesis focuses on describing the acoustic effects of emotion or, in other words, how speech is affected by the emotion of the speaker. One of the main reasons for this is the challenge of developing high quality human-machine interaction, where the machine would be able to recognize the emotions of the user and take actions accordingly, as well as interact with the user using highly intelligible and natural-sounding speech, even expressing human-like emotions appropriate to the situation (see section 2.2.2). But, as was discussed in section 2.4.2, conveying emotions or attitudes is one of the paralinguistic uses of prosody. It is therefore reasonably argued that naturalness highly depends on appropriate prosodic modelling and the essential inclusion of expression/emotion in synthesized speech (Schroeder 2004).

According to (Murray and Arnott 1993), vocal correlates of emotion (that can be used for synthesis) can be divided into five groups: Pitch-related features, formant frequencies, timing features, voice-

quality parameters and articulation parameters. However, emotional speech synthesis studies are often restricted to the prosodic parameters only (Schroeder 2004), although several other acoustic correlates have been studied as well (e.g. Xiao *et al.* 2005). One example is voice quality (Laver 1980), that has been studied as an acoustic correlate of emotion (Johnstone and Scherer 1999; Gobl *et al.* 2002), but is rarely treated as such in emotional speech synthesis studies (Schroeder 2004). The issue of naturalness in emotional speech synthesis revolves around two themes. The first of these relates to theoretical perspectives and definitions of human emotion, while the second is the issue of obtaining recorded speech which contains genuine emotions.

In the past, research in emotional speech syntesis had been based on traditional theoretical perspectives of emotion (Cornelius 2000), which describe "fully blown" emotional states, described by labels such as "anger", "fear", "disgust", "happiness", "sadness", "surprise" etc. However, it was argued (Cowie and Cornelius 2003) that these impressionistic descriptions of emotion are not on par with normal everyday-life speech, since emotional labels are ambiguous and subjective both in perception as well as attribution of a label to an utterance. Thus, alternative representations of emotions, such as that of the circumplex model (Russell 1997), became prominent. This perspective describes emotional continuums (rather than states); a continuum is defined by a number of perpendicular axis, or dimensions. The most prominent such dimensions are those of activation, evaluation and power (Schroeder *et al.* 2001). Correlations between positive/negative directions along these dimensions and several a/p features have been studied in (Schroeder 2004).

Therefore, both distinct emotional state approaches and dimensional models attempt to quantify a relationship between prosodic parameters and emotion (Schroeder *et al.* 2001).The validity of this approach depends on the speech material (corpora) that are available for analysis and, in particular, the validity of the emotional content in the recordings. The validity can be evaluated by recognition rates in listening experiments: If the intended emotions are perceived as such, then the reliability of the content can be considered satisfactory. (Campbell 2000) categorized various methods of acquisition of speech material used in studies of emotional speech.

a) *Acted speech*, where actors perform the intended emotions, produces the most recognizable results, but it is arguable that this is because actors are trained to exaggerate their emotional displays, so that they are easily recognized by their audience. It has been argued (Kehrein 2002) that this type of expression is very distant from the type of expression encountered in real-life conversations. The advantage of this method is that it can produce sentence pairs of content-neutral texts acted with different emotions, which can be used to model variations in a/p features due to emotion in a straightforward way. In (Banse and Schrerer 1996), a number

of actors spoke two pseudo-sentences (nonsense sentences), while performing 14 distinct emotional states. The recordings were rated by experts for recognizability and a further selection was made according to recognition rates during a number of listening tests. The selected recordings were analysed, particularly studying the variations of F0, speech rate, mean energy, and spectral features. After discarding all variations related to speaker gender, age, and identity, it was found that emotion is responsible for a large percentage of the variations.

b) *Context-based stimulation* refers to a procedure in which subjects are reading aloud a text that stimulates a specific emotion. The recognition rates in this case are high, but this could be because there are many linguistic cues in the recording that listeners can base their assessment on. A solution to this problem was suggested by (Campbell 2000): recognition rates can be obtained from subjective tests that are based on re-synthesis of the acquired prosodic contours on content-neutral sentences, thus removing the linguistic cues.

c) *Found* speech corpora (from radio/TV shows, broadcast news etc) are also used in studies of emotional speech, but the argument remains that newscasters and people generally in a studio are still "performing", rather than displaying their natural, everyday emotional code. Other "found" recordings, such as radio transmissions during dramatic situations (such as the Hindenburg crash radio broadcast[7]) or recordings in public places may overcome this problem, but the audio quality of such recordings is often inadequate. Recorded telephone conversations from customer care services is another example of large corpora with high emotive content (such as customers expressing their dissatisfaction with a product/service) but legal issues with releasing such material often become a barrier to their use for research.

d) Finally, *emotion elicitation* makes use of mood induction procedures (MIPs) (Gerrards-Hesse *et al.* 1994), which are experiments designed to induce emotive reactions to the subjects. For example, (Johnstone 1996) used computer games to induce emotional stimuli to the subjects, who had to report on their progress in the game verbally. The main advantage of this method is that it produces recordings of *spontaneous* speech, which can be argued to contain the most genuine emotions that is possible to record in a laboratory environment. There are ethical issues to consider when designing such experiments. For example, it is unethical to induce negative emotions to the subjects. Another disadvantage is that it is

7 One of the finest German passenger zeppelins, the Hindenburg, crashed on May 6th, 1937 in Lakehurst, New Jersey, while attempting a mooring. Engineer C. Nehlsen was recording the mooring process. When the disaster happened, Nehlsen continued describing the events as they occurred, thus producing an "emotional" recording, in which his expression clearly shows he is shocked and overcome by the tragedy. [online: http://www.otr.com/hindenburg.html, (01/04/2010)]

difficult to build a large corpus, as such a process requires a significant amount of time and resources. However, some significant work in this direction has been reported (Maekawa *et al.* 2000; Cullen 2008a).

2.4.4 Prosody in Spoken Dialogue systems

In the previous sections, research on several functions of a/p features has been presented. However, most of that work has been focused on monologue speech (Macchi 1998; Campbell 2006), essentially neglecting prosodic functions in relation to dialogue (Kohler 2004), although it has been known for quite some time that there are significant differences in prosody of monologue and dialogue speech (e.g. Hirose *et al.* 1996). There are exceptions to this research tradition (e.g. Bruce *et al.* 1996) and, recently, prosody has been taken into account in the context of research in SDS. Due to the multi-functional role that prosody holds in spoken communication, there exist many different studies on how prosody can be utilized to improve performance of SDS. (Swerts and Terken 2002) distinguishes three main themes:

(a) Improving performance of the ASR component: Prosody can help re-segment previously ill-segmented utterances and improve the overall recognition rate, therefore reducing recognition errors and the need for additional clarification prompts from the system.

(b) Improving performance of the synthesizer, by utilizing *utterance generation*, in other words formulating an utterance that can be delivered with an appropriate prosody, providing for smoother and more pleasant dialogue.

(c) Interaction management which relies on dialogue act classification (see section 2.3.2), can utilize prosody as one of its classifiers (see below).

Of the above, (c) especially above is attracting a lot of interest in the research community since the review in (Swerts and Terken 2002). Prosodic information is usually combined with lexical and semantic information in order to improve the performance of dialog-act tagging (Hastie *et al.* 2002; Cerrato 2002; Ang *et al.* 2005; Rangarajan *et al.* 2007). (Rangarajan *et al.* 2007) reported a 74% accuracy rate using prosodic and acoustic cues only, compared to a 9% increase when combining lexical information and only marginal improvement when combining syntactic information and syntax-based prosody. As mentioned in section 2.3.2, the classification of dialogue acts is crucial for implementing sophisticated interaction (turn-taking) in dialogue systems (Raux 2008). According to (Lennes and Anttila 2002):

"Turn-taking dynamics are related to systematic changes in the prosodic and acoustic properties of speech, but such processes are not well understood."

The latter study found significant correlation between low-level acoustic and prosodic features (overall time-share, F0, tempo, pauses, creakiness) and turn switches or topic changes in dialogues in Finnish. Moreover, (Lennes and Anttila 2002) identified differences in these patterns across languages (namely between Finnish and English). (Edlund *et al.* 2005) also implemented an utterance segmentation and turn-taking methodology for dialogue speech based on online prosodic analysis.

In addition, the issue of emotional speech, which is also related to a/p features (see section 2.4.3) has come into attention in the context of dialogue speech, not only as a possible improvement of naturalness, that can arguably be accomplished by synthesizing a system voice that conveys appropriate emotional/attitudinal behaviour, but also as a method of detecting user emotions during human-computer interaction. Although user emotions can be relevant in various tasks (e.g. Fernandez and Picard 2000), a direct application can be the detection of user frustration, which can lead to better error-detection (Holzapfel *et al.* 2002; Lee and Narayanan 2005; Austermann *et al.* 2005).

Therefore, prosody has been identified as a major avenue of improving the naturalness or human-likeness of SDS both in recognizing user emotions and synthesizing expressive speech, as well as re-defining prosodic modeling (utterance generation and dialogue act-tagging) in a dialogue context. Dialogue act classification and emotion recoginition can benefit from a quantitative description of inter-speaker accommodation of a/p features, as the latter features are prominent classifiers in these techniques. Similarly, utterance generation can be improved significantly by implementing a/p feature accommodation in the prosodic model of the synthesis component. However, as discussed in section 2.2.4, an analytical study of inter-speaker accommodation requires the acquisition of a corpus of natural human dialogues. This issue is discussed in the next section.

2.5 Recordings of natural speech

As mentioned earlier (section 2.2.3), the acquisition of natural human speech recordings is a major issue in speech technology in general. In section 2.4.3, this issue was discussed in relation to the naturalness of the *emotional* content of the utterances in the corpus. However, the requirement for recordings of natural speech is not restricted to emotional speech synthesis and recognition. A study of inter-speaker accommodation has to be based on such recordings as well, as discussed in section 2.2.4: human dialogues are the target against which the naturalness of human-machine interaction can be evaluated.

There exist two prominent sources of natural speech recordings. In-lab experiments and real data,

which correspond roughly to “found” and “elicited” speech corpora, respectively, according to the terminology of (Campbell 2000) which was presented in section 2.4.3. There are three criteria for evaluating methods of speech corpus acquisition that are implied in that description: feasibility/resource cost, audio quality and naturalness of the content. A fourth criterion is *re-useability*: if the content can be re-used for several research studies, then the resource cost is balanced. The difference between re-usability and resource cost is that the former is a property of the final corpus (as are audio quality and naturalness) whereas resource cost is only considered before and during the process of acquiring the recordings.

Audio quality is an issue commonly overlooked in speech technology, although the advent of high-throughput computers and state-of-the-art audio equipment has minimized this problem. However, there are issues with the currently available corpora. For example, found speech corpora of telephone conversations are the largest currently available (Furui *et al.* 2005), but the quality of the recordings is questionable. Telephone quality is typically of an 8KHz sampling rate combined with low-pass filtering, a specification that is unacceptable both in comparison to modern audio processing standards, as well as because it effectively omits a large band of frequencies that falls within the audible range. According to (Katz 2002), even CD quality (44KHz/16-bit), which is often considered as “top-level”, is in reality a minimum standard in state-of-the-art audio recording and production. Noise contamination can also affect recordings not carried out in a laboratory environment, for example when a microphone is installed in a bus or an underground rail car (Campbell 2000), or due to distortion introduced by compression when the voice signal is transmitted over telephone.

Perhaps the most significant criterion, however, is that of naturalness of the recorded speech. The term most commonly used to describe real-life occurring speech, is *spontaneous* speech, as opposed to speech that is read from text or acted or performed in any other way that is planned in advance (Stolcke *et al.* 1998). As pointed out in (Furui *et al.* 2005):

> “Both acoustically and linguistically, spontaneous speech and speech read from a text are very different. Spontaneous speech includes filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies”

In order to overcome the audio quality problems with *found* speech mentioned above, there is the solution of recording spontaneous speech in a laboratory environment. However, people tend to “perform” when put in front of a microphone, and in some cases they become anxious. It is difficult to inspire a relaxed atmosphere during the recording session, due to the presence of the audio equipment. Thus, the task of collecting genuine spontaneous speech of laboratory quality (noise free

and high sampling rate) is difficult.

A compromising solution arises from the psychological studies on mood induction procedures (MIPS) (Gerrards-Hesse *et al.* 1994) and, in particular, *task-based* MIPS. These generally employ a scenario where isolated subjects are asked to perform seemingly simple tasks while their speech is being recorded. In (Kehrein 2002), the task given was assembling a LEGO puzzle. One subject provided instructions from the manual while another was trying to assemble the pieces together. Artificial nuances (such as missing pieces) were used to stimulate expressive responses from the speakers. In (Johnstone *et al.* 2005), speakers were recorded while playing computer games, with events in the game providing the necessary stimuli for expressive, spontaneous speech. Although both examples above were employed in studies of emotional speech, task-based experiments are relevant to recording spontaneous dialogue in general (e.g. Bomsdorf and Szwillus 1999). Despite the fact that the speakers are aware that they are being recorded, an artificially created task-based situation can provide the necessary context to help diminish the effect of that awareness. Thus, task-based scenarios have been used in order to record "natural" dialogue as, for example, in (Kurematsu *et al.* 2000), where subjects were asked to collaborate in making travel and accommodation arrangements, based on conflicting schedules and timetables provided.

In conclusion, carefully designed task-based experiments are the most appropriate means for recording natural, spontaneous dialogues in a laboratory environment, although existing recordings of dialogues from real applications (such as customer assistance call-centers), can also be used if the audio quality is acceptable.

2.6 Discussion

Spoken dialogue systems have reached a point at which the goal of human-like conversation is being considered both as a means of improving on the naturalness of the interaction, as well as a means of increasing efficiency and making possible the extension of the application field to more complex tasks. The former is a long-term goal of speech technology in general, as indicated by the literature review on prosodic modeling, emotional speech and, the more recent turn towards conversational spoken dialogue systems that exhibit human-like conversational capabilities. The second goal (increased efficiency) is driven by observations on the efficiency of human dialogues in problem solving and reasoning tasks, and the inadequacy of current SDS to deal with these complex domains. There is a clear distinction between these two lines of research, as human-likeness is not directly related to usability (Pieraccini and Huerta 2005). The two goals are followed through distinct (but parallel) lines of research .

Human-likeness, which can refer to any property of human-computer interaction which resembles human dialogue, is sought after through research on human dialogue corpora, as was discussed in section 2.2.4. Within this framework, the content of the corpus is crucial to characterizing human speech phenomena. As discussed in section 2.5, the most credible source of natural human interaction are corpora of spontaneous dialogues (either found or elicited). The importance of using dialogue recordings is evident from studies that showed the inadequacy of monologue-based methods to characterize the variable properties of speech in various domains of research, such as emotional speech (Batliner *et al.* 2000) and prosodic modeling (Hirose *et al.* 1996). Off-line analysis of human dialogues leads to models of human-human interaction which can possibly *guide* design principles for SDS (Larsson 2005), but are often incompatible to industry standards or even more complex architectures that are only implemented in the research domain (Dybkjær and Dybkjær 2004). However, spoken dialogue research is the primary path for improving naturalness of SDS (Larsson 2005), as human dialogues are the only evaluation targets for assessing the perception of human-likeness (Edlund *et al.* 2008).

(Edlund *et al.* 2008) points out that human-like interaction is not suitable for all tasks and, in some cases, it may actually *hinder* efficiency (Pieraccini and Huerta 2005). In addition, it has been argued that, although speech is a natural and efficient way of communication, it may not always be the most suitable (Larsson 2005). In some cases, a GUI, or a combination of a GUI and an SDS can be much more efficient (e.g. in city guides). In addition, (Edlund *et al.* 2008) points out another possible pitfall, namely the "uncanny valley": a system that is too human-like, so that it feels awkward and causes dis-comfort to the user community. The answer of (Edlund *et al.* 2008) to this is that, given the current capabilities of commercial SDS, it is premature to think about this problem.

Conversely, efficiency has traditionally been accomplished by constraining the interaction and choosing sufficiently limited domains. However, human-likeness is desirable for many commercial SDS applications, as it would increase pleasantness and user satisfaction, which is also a significant benchmark in the SDS industry (Moller *et al.* 2007). In addition, implementation of certain aspects of human interaction, such as the collaborative nature of dialogue (Traum and Allen 1994), is desirable in order to extend the application field of SDS into more complex tasks, such as problem solving (Dybkjær and Dybkjær 2004). The main focus of research in this area is put on characterization of the discourse structure (Mushin *et al.* 2003), in order to allow SDS to manage dialogues more *efficiently.*

Importantly, prosody plays a key role in both of the above lines of research. For example, an

improvement in naturalness of SDS is the implementation of prosodic models suitable for dialogue speech (e.g. Hirose *et al.* 1996), in contrast to traditional prosodic models that have been based on well-formed monologue sentences (Kohler 2004). Discourse structure and dialogue speech segmentation also depend on classification of prosodic features (Bruce *et al.* 1996). Emotion recognition and emotional speech synthesis are also based on prosodic features (Lee and Narayanan 2005) and are simultaneously utilized in utterance generation (naturalness) and error-detection (efficiency). Therefore, it is likely that inter-speaker accommodation of prosodic features, if implemented in SDS, will improve human-likeness, by simulating this behaviour, as well as efficiency, by informing dialogue act and emotion classification with output from online monitoring of prosodic accommodation.

Another significant issue in human-computer dialogues is that of interaction management in terms of the *temporal* organization of the interaction (inter-speaker silence duration and occurrence of overlapping speech). Again, two lines of research can be distinguished here. On one hand, the functional description of turn-taking and back-channeling feedback cues aims to identify methods for SDS to take or release the floor in a way that reduces latencies (Raux 2008) and allows for user "barge-ins" (e.g. Glass 1999). This line of research builds upon current half-duplex representations of dialogue and approaches human-like conversation incrementally upwards. On the other hand, research on human dialogues has indicated coupling of interlocutors in closed-loop systems that exhibit synchrony, feedback and simultaneous activity, (Campbell 2009; Heylen 2009). While theories on rigid coupling of rhythm (e.g. Wilson and Wilson 2005) have not sufficiently captured the temporal accommodation of turn taking (Benus 2009), there is significant evidence of temporal accommodation in human dialogues (Bosch *et al.* 2005). This is particularly the case in spontaneous dialogue speech. These findings provide further motivation for investigating accommodation phenomena in spontaneous human dialogues.

(Edlund *et al.* 2008) proposed a complete framework of implementing and evaluating human-like behaviour in spoken dialogue systems. This framework suggests feature extraction from recordings of human dialogues (in order to formulate a description of a particular phenomenon, such as a simple model), and a range of alternative evaluation methods for implementing similar behaviour in SDS. In this case, the evaluation target is not necessarily the *perceived* naturalness (the usual case in monologue speech tradition), but the *similarity* of the human-machine manifestation of the investigated phenomena to the human dialogue manifestation. A further distinction is made between evaluating whether the system voice resembles that of a human in some aspect of dialogue, and/or the user responds to the system similarly to a human in a human dialogue. The former evaluates the

feasibility (or goodness) of the implementation, while the latter tests the user *response* to the modeled behaviour. (Edlund *et al.* 2008) also suggested that incorporation of inter-speaker accommodation phenomena may improve SDS, both in pleasantness and efficiency, as accommodation is known to have a communicative as well as a social function.

In conclusion, the background review in this chapter has identified the study of inter-speaker accommodation as a promising route towards improving SDS in a number of ways. The following chapter presents a review of inter-speaker accommodation phenomena in human dialogues.

3 Inter-speaker accommodation in human interaction

3.1 Overview

In the previous chapter, inter-speaker accommodation was identified as a property of human interaction that can improve current SDS primarily in terms of naturalness but also in terms of efficiency. This chapter presents a review of theoretical studies on accommodation in human dialogues that are primarily focused on its function in human interaction. An understanding of the function of accommodation is required in order to inform SDS design, in terms of simulating this type of behaviour in a way that serves a similar function.

The basic concept of inter-speaker accommodation is that two (or more) individuals engaging in dialogue tend to show similar behaviour in respect to various aspects of their speech; prosody, accent, lexical and syntactic choice, as well as temporal features which involve turn-taking behaviour, such as the duration of intra-speaker and inter-speaker pauses and the occurrence of overlapping speech; and this behavioural "adaptation" extends to other modalities in face-to-face to conversation; gestural and postural behaviour of one matches or complements that of another while engaged in dialogue. This phenomenon is generally believed to be ubiquitous, and -most of the time- unnoticed, at least at the higher levels of consciousness, but can also be an intended strategy with specific communication goals.

Apart from evidence presented here and elsewhere, this phenomenon is intuitively known in general: one common example is well known to people who have grown up in a region with a characteristic regional accent but have moved elsewhere, for example to a big city. These people typically adopt a more neutral and widely accepted accent in their everyday city life, but can readily switch back to their regional accent as soon as they return to their home region, even without consciously deciding to do so. Another typically occurring situation is when fluent, native speakers of any language match a non-native (and less fluent) interlocutor's erroneous grammatical/syntactic forms, as they believe this to make the communication more efficient. The latter is an example of a conscious choice to adapt one's speech.

A basic distinction that has to be drawn, is that between studies on inter-speaker accommodation which are discussed here, and studies on the *collaborative nature of dialogue*, which lie in the field of conversation and discourse analysis and were discussed (briefly) in the previous chapter (see section 2.3.2). Accommodation phenomena are mostly studied in psychology and psycholinguistics, communication science, and cognitive sciences. In addition, they have been studied in human-computer interaction, even before the emergence of SDS. For example, it has been reported that

users adapted their lexical choices to those of a text-based interface (Zoltan-Ford 1991). In addition, interest in accommodation phenomena has been recently refueled in the context of SDS (Edlund *et al.* 2008).

(Burgoon *et al.* 1995) describes a variety of behavioral patterns emerging in both verbal and non-verbal communication: adaptive responses, accommodation, convergence, matching, mimicry, synchrony, reciprocity, complementarity. All of the above observed interaction behavioral patterns are "non-random, patterned, or synchronized in both timing and form" (Bernieri and Rosenthal 1991). Importantly, the patterns exhibited by two interactants are similar or dissimilar in form, as in the case of divergence, dis-synchrony, non-accommodation etc.

Due to the diversity of approaches arising from the different - but relevant - fields of research that were mentioned in the second paragraph, there is also diversity in terminology, definitions, research goals and methods used. Such situations allow for several categorizations of the studies found in the literature, of which there exists a multitude. Given the lack of universally adopted definitions, this text will use "inter-speaker accommodation" or simply "accommodation" to collectively describe any of the phenomena described in this chapter. It is also noted that the theories described here refer to interpersonal behaviour in general, which is not restricted to speech, but also includes other modalities such as body movement and posture, hand and facial gestures, gaze and eye movement.

3.2 Terminology and definitions

A number of terms have come to prominence over decades of research on inter-speaker accommodation: convergence, accommodation (Giles *et al.* 1987); coordination, inter-speaker influence (Jaffe and Feldstein 1970; 2001); alignment (Pickering and Garrod 2004); entrainment (Brennan 1996); behavioural matching, adaptation (Burgoon *et al.* 1995); synchrony tendency (Nagaoka *et al.* 2005); and synchrony (Campbell 2009). There is some confusion arising from the multitude of terms and the fact that they are often used under different definitions. Moreover, although they all fall under the same basic concept that was described in the previous section, there are subtle differences that are often overlooked. According to (Warner 2002), one of the most consistent terminologies is that of (Burgoon *et al.* 1995):

Behavioral matching is an 'umbrella-term', introduced to summarize many of the above observed behaviors. It refers to greatly similar or even *identical* patterns of behavior, between two or more interactants.

Complementarity is the opposite of matching, and refers to dissimilar behaviors that complement each other.

Convergence is the process by which the observed behaviors of two interactants, although dissimilar at the start of the interaction, are moving towards behavioral matching.

Divergence is the opposite of convergence and refers to the behavior of moving towards a dissimilar pattern, therefore indicating a change of behavior for at least one of the interactants.

Mirroring involves visual behaviors (such as posture) and refers to the interactants keeping an identical posture or gaze.

Synchrony is a temporal equivalent of mirroring, in that it refers to similar or identical rhythmic/temporal patterns exhibited by the interactants.

Reciprocity is the tendency to respond positively to the interaction by exhibiting a similar behavior and, according to Burgoon et al, is reflected by both mirroring and synchrony.

Dissynchrony is the opposite of synchrony and, as implied by its name, refers to the interactants exhibiting non-synchronous temporal or rhythmic behaviors.

Compensation in a narrow sense is the opposite of mirroring, as in keeping one's gaze or posture opposite to that of another, but in a broader sense implies a behavior opposite to reciprocity: avoidance of matching expectations, adopting behavioral patterns towards opposite directions.

However, as pointed out in the previous section, but also elsewhere (Warner 2002; Edlund *et al.* 2009), there are no universally adopted definitions. In the seminal presentation of *Speech Accommodation Theory* (Giles *et al.* 1987), convergence is defined as "a linguistic strategy whereby individuals adapt to each other's speech by means of a wide range of linguistic strategies, including speech rates, pauses and utterance length, pronunciations and so on.". Hence, there is no mention of an *evolving process* in this definition, in contrast to the definition in (Burgoon *et al.* 1995), although the theory itself implies it. (Warner 2002) reports that synchrony has been used for at least two distinct measurements: "global observer judgments" and "synchronized cycles detected by cross-spectral analysis". (Edlund *et al.* 2009) adopted standard dictionary definitions, arriving at a definition for convergence virtually identical to that of (Burgoon *et al.* 1995) (movement from initial dissimilarity towards similarity), but a different one for synchrony: "... phenomena that happen at the same time or work at the same speed". There is a referential mismatch here, in that (Edlund *et al.* 2009) refers to "synchronous phenomena", the *contemporaneous,* or *synchronous* variation of any feature of the speech signals of two interactants, whereas the definition of (Burgoon *et al.* 1995) only refers to *temporal* (duration, latency) aspects of speech.

As will become clearer in chapter 4, the operational definition adopted in each case suits the proposed methodology or theory, hence the diversity of definitions. Rather than adopting any of these definitions, each is selectively conceptualized with respect to its proponents' theory or

methodology. In the next sections, the most representative such theories are presented.

3.3 Perspectives of inter-speaker accommodation

Inter-speaker accommodation phenomena have been described by several theories (or models), of which a comprehensive review can be found in (Burgoon *et al.* 1995). In that review, the models are categorized along a "continuum", in which four basic categories of models are identified, as shown in Table 3.1. At one end of the continuum, there are the physiological and biological models that consider accommodation phenomena as automatic, non-conscious reactions. From a biological point of view, convergence and synchronization is seen as advantageous to survival, as well as a sign of intimacy. This point of view is supported by observations of similar behaviour exhibited (in non-verbal communication) by other species (Oviatt *et al.* 2004). Accommodation at this level is also seen as serving communication efficiency (Giles *et al.* 1987; Pickering and Garrod 2004).

The term *interactional synchrony* was first proposed by Condon and Ogston (1966, 1967; 1971) as a means of describing listener body movements as affected by speech and movements of the speaker. The phenomenon, which was considered as a non-conscious autonomous behaviour, was later observed on mother-infant interactions (Gratier 2003) and was also related to infant development (Jaffe *et al.* 2001).

Motor mimicry, also termed *mirroring*, refers to mimicking (or mirroring) an emotive expression of another person, and had already been observed by Adam Smith, Herbert Spencer, and Charles Darwin (Bavelas *et al.* 1986). One example of motor mimicry is people watching an accident scene on video and making a "painful" facial gesture. Traditional accounts of motor mimicry attributed this behaviour to the *individual,* as "a primitive empathy", a trait (empathic ability), a means of expressing a vicarious emotion, or a signal of "taking the role of the other" (Bavelas *et al.* 1986).

A second category (arousal and affect models) consists of theories that, in addition to the above biological needs, propose that matching behavioural patterns in interactions satisfy psychological needs. According to the *Affiliative Conflict Theory* (Argyle and Dean 1965), human interaction is characterized by an equilibrium of immediacy (or intimacy). If there is an action by one of the partners that causes the interaction to deviate from the equilibrium point, this causes anxiety (or arousal) to the other partner, who tries to re-establish the equilibrium by means of *compensation* (see definition in previous section). The *Arousal-Labeling Theory* (Patterson 1976) introduced reciprocity into this description, by positing that departures from the equilibrium that are large enough to cause arousal are "labeled" positively or negatively, thus producing reciprocal or compensatory behavioural patterns, respectively. A further expansion was introduced by the *Bidimensional model* (Kaplan and Kaplan 1984), which considered a two-dimensional approach:

manifestations of reciprocity and compensation are caused by the psychological needs of intimacy *and* social control.

Reactive, automatic, non-symbolic, indicative behaviour		
↑	**I. BIOLOGICAL MODELS** (based on comfort needs, safety, bonding, social organization, universal processes) **Interactional Synchrony** (Condon and Ogston 1966) **Motor Mimicry and Mirroring** (cf Bavelas *et al.* 1986)	Biological and Psychological Needs – Focus on Individual
↑	**II. AROUSAL AND AFFECT MODELS** (addition of psychological needs to above factors) **Affiliative Conflict Theory** (Argyle and Dean 1965) **Arousal-Labeling Theory** (Patterson 1976) **Bidimensional Model** (Kaplan and Kaplan 1984) **Discrepancy-Arousal Theory** (Cappella and Green 1982) **Dialectical Models** (Altman *et al.* 1981)	Biological and Psychological Needs – Focus on Individual
Habitual Behaviour		
↓	**III. SOCIAL NORM MODELS** (incorporation of cultural, societal factors, ingroup-outgroup relations) **Norm of Reciprocity** (Gouldner 1960) **Social Exchange and Resource Exchange Theories** (Homans 1958) **The Dyadic Effect** (Jourard and Landsman 1960) **Communication Accommodation Theory** (Giles *et al.* 1987)	Social Processes, Societal Needs – Focus on Groups
↓	**IV. COMMUNICATION AND COGNITION MODELS** (emphasis on functions, goals, meanings, perceptions, attributions) **Sequential-Functional Model** (Patterson 1982) **Expectancy Violations Theory** (Burgoon 1978) **Cognitive-Valence Theory** (Andersen 1999) **Motor Mimicry Revisited** (Bavelas *et al.* 1986)	Hybrid Needs and Goals – Focus on Dyads
Communication, mindful, intentional, symbolic		

Table 3.1: Categorization of interactional theories, adapted from (Burgoon et al 1995)

Discrepancy Arousal Theory (Cappella and Green 1982) proposed that arousal occurs from discrepancies from expected behaviour of an interlocutor. These expectations are based on familiarity, acquaintance and an established level of intimacy. Again, the discrepancies can be evaluated positively or negatively, giving rise to reciprocal or compensatory responses. *Dialectical models* (Altman *et al.* 1981) consider psychological needs as resulting from cyclic fluctuations (oscillations) driven by oppositional forces which occur in everyday interaction (rather than being biologically based as in the previous theories). These oppositional forces are various: autonomy vs connection, openness vs closeness and novelty vs predictability. Dialectical tension, in contrast to discrepancy or equilibrium violation, is seen as neither good or bad; thus, the theory predicts that interactants may reciprocate or compensate, depending on whether they attempt to match their

psychological needs (such as in stable relationships).

The third category, labeled "social norm models" differs from the other two in that, instead of focusing on the individual, focuses more on social relationships. The principle of s*imilarity attraction* is one example of this: in order to become more attractive, people attempt to appear similar (or converge) to their attraction targets. Thus, people try to look similar to others, in order to be liked, or accepted, and adapting their speech to that of others is one way to express this similarity (Giles *et al.* 1987).

The *norm of reciprocity* (Gouldner 1960) refers to the expectation that people tend to respond positively to positive action or attitude towards them, and negatively or indifferently towards negative or harmful action/attitude. This is a social principle (or norm), which derives from the need for survival, as it encourages cooperation in order to survive in hostile conditions (Aronson 2007). *Social Exchange Theory* (Homans 1958) posits an economic model for human social relationships, in that people's behaviour can perhaps be explained on the basis of a subjective "cost-benefit" analysis. The *Dyadic Effect* (Jourard and Landsman 1960) relates to the degree of self-disclosure in dyadic relationships, as it has been observed that interpersonal feedback elicits the same from others. All of the above social norms have been related to behavioral matching in some way (Burgoon *et al.* 1995). Perhaps the most representative and popular theory from this category is the *Communication Accommodation Theory* (Giles *et al.* 1987; 1992) which is explained in more detail in section 3.4.

The opposite end of the categorization continuum, the *communication and cognition* models, groups together those theories that describe interlocutor similarity phenomena as conscious or intentional from the point of view of the interactants who are typically well-acquainted *dyads* (e.g. married couples). Well-acquainted interactants usually develop communication "norms" - which both adhere to – over time, and departure from that norm by either speaker violates the expectations of the other, thus giving a warning sign that the situation requires attention. A description of the *relationship* between the interactants, as well as their goals and expectations, is central to these theories, since this knowledge is required in order to explain the interaction phenomena in this way. The *Expectancy Violations Theory* (Burgoon 1978) bears many similarities to the Discrepancy Arousal Theory mentioned above. One of the major differences is that in the former, the expectations are not limited to arousal and affect but are formed through acquaintance, are known as social norms, or they are specific to a particular interaction. *Cognitive Valence Theory* (Andersen 1999) similarly proposes six schemata (culture, personality traits, state, situation, interpersonal valence, relationship), in order to explain the valence (positive or negative) of perception of intimacy behaviour by either partner in a dyadic or social relationship. Finally, the revision of motor

mimicry (Bavelas *et al.* 1986) takes the focus away from the individual and proposes that this phenomenon has a social function of empathy and indication of similarity towards a conversational partner.

The categorization of (Burgoon *et al.* 1995), although informative and wide in scope, should not be followed strictly; there is overlap between the categories, hence the description of the categorization as a "continuum" rather than a categorical classification. Communication Accommodation Theory (CAT) (Giles *et al.* 1987), for example, describes the phenomena as both autonomous as well as intended behaviour. The classification of a particular theory into one of the four categories is a good indication of its main focus at best. This is particularly the case for *Interpersonal Adaptation Theory* (IAT), which is proposed in the same text as the review that is summarized here (Burgoon *et al.* 1995), as it lends from all four categories. The same applies to the Interactive Alignment Model (Pickering and Garrod 2004), which describes "alignment" between interactants during dialogues at various levels, from non-verbal low-level signal features to lexical/syntactic and further on to semantic/symbolic representations and situational models, covering the entire range from spontaneous adaptation to conscious actions. Nevertheless, the review of (Burgoon *et al.* 1995) serves well as a starting point to understanding the scope of functions attributed to inter-speaker accommodation phenomena.

3.4 Communication Accommodation Theory

The phenomenon of speech accommodation in dialogues has been studied and introduced into the framework of *Speech Accomodation Theory* (SAT) over two decades ago (Giles *et al.* 1987). This framework, that was later renamed *Communication Accomodation Theory* (CAT), proposes that accommodation of speech features (accent, speed, pause duration, lexical) occurs as a communication strategy (either conscious or unconscious), with specific social goals (integration into a social group, or identification with a member of the same group). In this section, a summary of the main ideas of SAT (and CAT) is given.

3.4.1 Convergence and divergence

SAT defined *convergence* as a linguistic strategy. Convergence refers to adaptation of an individual's speech characteristics (pause duration, speech rate, utterance length, accent, etc.) in order to match those of a partner in dialogue more closely. Similarly, *divergence* refers to a tendency of the individual to maintain their distinct speech style by accentuating differences in the aforementioned characteristics of speech. SAT distinguishes between *upward* and *downward*

convergence (or divergence), the former refering to changing one's style in order to match a valued social status profile, while the latter suggesting 'shifting' towards less valued social profiles, such as a language variant specific to ethnic/cultural/social groups or non-fluent speech/illiteracy.

Futher, SAT proposes further categorizations of convergence (and divergence) by relative movements or 'shifts' between two interlocutors (A,B). Thus convergence or divergence can be mutual (A → ← B, ← A B →) or non-mutual (A → B, A B →), or one speaker might converge while the partner diverges (A→ B→).

Another distinction is introduced by the difference that lies between a *manifest* speech style and the *perception* of that speech style that is biased by a stereotypical belief. (Giles *et al.* 1987) pointed out that both convergence (or divergence) and its *evaluation* (how positively or negatively it is perceived) depend on one's *perception* of the other's speech, rather than their actual, *manifest* speech styles. A common example is imitation of a language variant by non-native speakers (such as a Dublin accent in Irish English): although characteristics of that accent might be prominent in the native speaker's manifest speech style, they might be perceived as *accentuated* by the non-native speaker, therefore misleading them to converge towards a similar speech style. From the point of view of the native speaker, that might be perceived as mocking of their social group, or as a comical social integration attempt at best.

Convergence (or divergence) can be additionaly distinguished into total and partial. The former refers to near absolute matching of speech style metrics, e.g. two interlocutors exhibiting very similar speech rate. The latter signifies a clear *movement* towards matching, such as increasing one's speech rate in order to converge to a higher rate of the interlocutor, but not to the extent of matching that speech rate.

Finally, convergence and divergence can be either *unimodal*, when accomodation occurs along only one characteristic of speech (such as speech rate, or accent), or *multimodal*, when two or more speech characteristics converge.

The central idea of SAT is that convergence (and divergence) is a strategy that humans engaging in dialogue use (either consciously or unconsciously) in order to achieve specific goals. In the landmark study of SAT (Giles *et al.* 1987), three such goals are proposed: social approval by the listener, serving communicational efficiency, and maintaining a positive social identity.

3.4.2 Communicative function of convergence

Within the framework of SAT, convergence is regarded as a readily available strategy, which is

utilized to invoke *similarity attraction* by the listener. The latter is a sociological principle (Giles *et al.* 1987), which states that attraction is more likely to occur towards individuals that display similarity in behavior. Reduction of dissimilarities of dialects, convergence of native speakers towards grammatical errors of non-fluent speakers, interviewees adjusting their speech to match the style of the interviewer, and sales people matching speech styles of their customers, are only a few examples of such cases given in (Giles *et al.* 1987).

According to SAT, convergence will not always be the best communication strategy, as its effect is moderated by *'situational constraints'*. Such constraints are introduced by *'sociolinguistic norms'* or, in simpler terms, what people believe 'is right' in a given situation. An example given in (Giles *et al.* 1987) is that of interviews in Australian English, were interviewees were rated higher if they were using a 'refined' rather than a 'broad' accent, regardless of the accent of the interviewer (who was switching accents between interviews). In this case, therefore, convergence of the interviewee towards the interviewer's accent was not rated favorably. Also, "powerful" speech style was more often rated favorably as a response to a "powerless" speech style, whereas convergence to a 'powerless' speech style was more often negatively evaluated.

SAT further advocates that the evaluation of convergence of the interlocutor to an individual's speech pattern is largely dependant on *causal attribution.* Listeners tend to evaluate the effort on the part of a converging speaker favorably, when they attribute that effort to the speaker's desire for social integration and attraction. When convergence is forced by situational constraints, it is rated less favorably. As pointed out in (Giles *et al.* 1987), although SAT defines speech accommodation as a *strategy*, that does not necessarily mean that it is a *conscious* one. Rather, (Giles *et al.* 1987) points to evidence of spontaneity and autonomy for speech accommodation at various cognitive levels. SAT advocates that speech accommodation may well be "*scripted"* behavior (established behavioural routines) in many cases, but one can be simultaneously making conscious decisions on the appropriate choice of speech style.

3.4.3 Communication function of divergence

Similarly to convergence, SAT proposes a number of communication goals for divergence: the main goal proposed is social identity maintenance, or the desire of individuals or groups to maintain a positive social identity, cultural pride and distinctiveness. A series of studies in the review of (Giles *et al.* 1987) provides many examples of ethnic minority group members accentuating their distinct dialects or accents when their ethnic identity is made more salient, or they encounter ethnically "threatening" situations. Gender is also proposed as a socially identifying factor as, in

one of the studies reviewed, men talking to women were found to sound "more masculine" when their gender was made more salient.

(Giles *et al.* 1987) points out that a distinction between non-convergence and divergence would be superficial or unnecessary at best. Non-convergence is a passive behavior towards the dialogue and the interlocutor, and its most extreme form of intended accentuation of distinct speech features has been termed as speech *dis-accommodation.* Divergence (by definition) means shifting one's speech style *away* from that of an interlocutor. According to (Giles *et al.* 1987), it is more likely that causal attribution plays a key role in the evaluation of divergence (similarly to convergence). After all, non-convergence may well be the result of repertoire constraints (as in the case of non-native speakers) or individual personality factors.

Another communicative function for divergence proposed by SAT is that of *cognitive organization*, i.e. to put the interaction (dialogue) in order, or to provide a 'mutual basis for communication'. A series of studies reviewed in (Giles *et al.* 1987) provides various examples of this communicative behavior: speakers who are unfamiliar with the host social group or the situational context, tend to accentuate their accent or employ other divergent strategies in order to indicate their unfamiliarity. The expected result of this is tolerance on the part of the host community members towards violations of situational norms on the part of the speaker.

Another example given is that of speakers diverging from a speech style that is uncomfortable for them, in order to encourage the interlocutor to converge to a different speech style, such as when talking slowly in an attempt to "cool down" a rapidly speaking interlocutor. Similar examples include therapy sessions, where clients may be invited to talk more when therapists talk less.

In certain situations, dissimilarities in the interlocutors' speech styles are expected, as is the case with interviewers and interviewees where the latter were more positively evaluated, when maintaining their 'refined' accent as opposed to converging *downwards* to the 'broad' accent exhibited by the former.

Finally, there are social norms that indicate a pattern of interaction where the interlocutors are expected to 'complement' each other's speech. This is more often made obvious in interactions between doctors and patients, teachers and pupils, parents and children and so on. As pointed out in (Giles *et al.* 1987), this complementary nature of speech patterns does not exclude the possibility of simultaneous convergence, along a different dimension (such as speech rate). The complexity (and multi-modality) of convergence and divergence are also highlighted in the text:

> "… it is not entirely impossible to concoct instances in which people may wish to converge,

diverge, and complement each other with regard to various verbal, vocal, and non-verbal forms *simultaneously*".

In conclusion, SAT provides a theoretical framework that attributes communicative functions to convergence and divergence of speech style. Speech style is used as a broad term and can denote anything from speech rate and pause duration to choice of words, utterance length, accent, dialect, and even switching languages (in bilingual communities). Of particular interest are the definitions for mutual/non-mutual, partial/total, unidirectional/bidirectional and unimodal/multimodal convergence (or divergence). Additionally, SAT proposes that a genuine effort to converge to another's speech style is likely to be evaluated positively, if the situational constraints do not suggest otherwise.

3.5 Interactive Alignment Model

A different approach to inter-speaker accommodation is given in (Pickering and Garrod 2004), which proposes the Interactive Alignment Model (IAM). IAM is in essence a cognitive theory focused on dialogue speech, in contrast to the former Autonomous Transmission account (Levelt 1983), which has been based on monologue speech. The remainder of this section summarizes the key points of IAM.

3.5.1 Alignment at different layers

IAM describes the process of alignment or, in other words, a matching of linguistic features among two interlocutors engaging in dialogue. The alignment occurs at different "layers" (phonetic, lexical) of dialogue communication. Thus, there is alignment at the lexical level (interlocutors matching their choice of words) and adoption of mutually agreed descriptive words or expressions, without any open agreement. In simpler terms, a word or expression that is used once to refer to a concept is later repeated by the interlocutor at an appropriate time. This is both an acknowledgment of comprehension and a sign of agreement on the use of that particular word from that point on in the conversation.

Further, there is alignment at the syntactic level, as the repetition of expressions, especially routinized ones, leads to utterances of similar or identical syntactic form. In addition, (Pickering and Garrod 2004) presents evidence that syntactic alignment can also occur in "complementary" form, when responses of either interactant complement the prompts of the other, both contributing to the formation of a single syntactic structure.

(Pickering and Garrod 2004) posits that there is also alignment at the semantic level, with interactants sharing the same situational model (a a multidimensional representation of a specific situation that is taking place). In the experiment described in (Pickering and Garrod 2004), this representation is spatial: two subjects are asked to identify the location of a dot on a square grid. Therefore, they tend to "define" a coordinate system, as this is the easiest way to achieve this task. This mutually agreed definition is not explicitly stated as such, but occurs through referential expressions such as "two from the bottom, one from the left". IAM posits that, although it is possible for the two interactants to have their own individual situational models (coordinate systems in this case) and rely on interpreting their partner's model while using their own to convey information, it is much more efficient for communication if they share the same one.

Alignment at the articulation level makes production and comprehension more efficient. Repeated words tend to become less well articulated, to the point that they are not easily recognized outside the dialogue context (if listened to in isolation). This also occurs when the repetition is produced by the partner, which implies that comprehension and production mechanisms of both interactants are aligned simultaneously. Finally, (Pickering and Garrod 2004) refers to previous findings of alignment in accent and speech rate, which they see as another layer (phonetic) in the proposed multilayer model.

3.5.2 Autonomous process

IAM claims that alignment at different layers occurs autonomously at each layer. The mechanism of this process is *priming*. Priming refers to influence over repetition of introduced signals (in this case speech), which are called *primes* (Kolb and Whishaw 2003). As an example, an utterance that introduces a representation, such as the spatial reference frame that was described above, is likely to act as a prime; either speaker is likely to adopt and re-use that representation in the course of the dialogue. According to the authors, there are different primes for the different layers of the model (a word for the lexical layer, a syntactic form for the syntactic level, etc). Thus, alignment occurs autonomously at each layer separately. However, IAM also argues that alignment at one layer leads to alignment at another layer.

Another driving mechanism for alignment is the principle of "parity of representations" between production and comprehension. This, according to (Pickering and Garrod 2004), is a controversial but widely adopted principle which states that a representation which has been acquired for comprehension can be used for production and vice versa. This explains, for example, why two interlocutors can complete each other's utterances.

Lastly, IAM suggests that there exist simple "repair mechanisms" in order to deal with misalignment. In the experiment described in (Pickering and Garrod 2004), there are occasions where subjects had adopted subtly different representations, thus requiring clarification at some point when it was realized that communication was inefficient. The interactive repair mechanism employs *grounding,* or in simpler terms, establishment of shared knowledge among interactants (Clark and Brennan 1991).

3.6 Discussion

Theoretical descriptions of inter-speaker accommodation phenomena have existed for a long time, and they have proposed several functions as explanations. A few key conclusions can be extracted from the review in this chapter.

The factors that affect accommodation are numerous: individual, biological, psychological, social, as well as situational and dyad-specific. From the point of view of SDS, these findings are interesting. A common misinterpretation in speech technology is that accommodation occurs automatically, while the potential use of other factors is largely ignored. For example, the fact that accommodation has a social function (as proposed by CAT) has important implications for SDS that use animated avatars and other such personifications. Perhaps an animated agent appears unsocial because his/her/its[8] speech does not converge to that of the user. Similarly, evolving interaction behaviours of well-acquainted dyads could be utilized in SDS for home use, which learn the behavioural patterns of the end-user and adapt towards them over many sessions (thus acquiring more training data). A suitable human metaphor (see section 2.2.2) for SDS in real applications is that of talking to a service providing agent. In this context, the talking agent should appear social, in which case at least the social aspect of accommodation, as described by SAT, should be taken into account.

In addition, accommodating behaviour may vary from autonomous, spontaneous, non-intended adaptation to semi-conscious or even deliberately implemented strategy. Several possibilities for SDS arise from these findings as well. Although the simplest design strategy would be the implementation of spontaneous convergence, one can imagine an SDS that articulates more clearly and reduces speech rate if a user is having problems or is taking too long to reply (e.g. elderly users). In addition, it is possible to elicit a specific speech style from the users (faster/slower, louder) by taking advantage of the fact that they themselves tend to align to a synthesized voice (see chapter 4). The purpose of this is to encourage a speech style which can be recognized more accurately by the ASR component.

8 'its' in case the talking agent is an animal character (e.g. Oviatt *et al.* 2004)

Although the theoretical descriptions of accommodation phenomena reviewed in this chapter provide several possibilities for improving human-likeness and performance in SDS, the data that they have been based on are not particularly useful for implementing similar behaviour in SDS. The majority of theories described in this chapter are based on empirical studies and experiments controlled in such a way as to provide evidence of the correlation of a function (social, dyadic, etc) and a particular accommodation phenomenon. A particular note has to be made on *longitudinal* studies, which monitor the behavior of dyads over long periods of time (often for several years). The motivation behind this method is that interaction patterns can be monitored as the relationship between the interactants develops. In addition, some of the studies that provided supporting evidence for the theories described in this chapter have been based on expert assessments of whether the interactants' behavioral patterns "match" or not, rather than direct measurements.

Therefore, despite the significant body of knowledge that has been acquired over decades of empirical research on inter-speaker accommodation phenomena, theoretical models have mostly focused on their function (biological, emotional, dyadic, social), while the form of their manifestation has not been adequately described in a way that is usable for SDS. In order for the possible improvements provided by the theoretical models to be explored in the context of SDS, a *quantification* of accommodation phenomena is required. The following chapter presents a review of studies which have proposed methodologies of measuring accommodation.

4 Measuring accommodation

4.1 Overview

In this chapter, previous studies on measurements of accommodation phenomena are reviewed. These studies come from a variety of research areas (such as psycholinguistics, human-computer interaction and speech technology) and are thus significantly dissimilar in their specific aims, objectives, and methods. However, they share the goal of quantifying the accommodation phenomena described in the previous chapter. There are four dimensions along which these studies can be categorized or described:

(a) The overall goal of the study: depending on the research area of the study, the goal can be investigation and characterization of phenomena in order to validate a theoretical hypothesis (e.g. functional relationship between accommodation and positive evaluation, or efficiency of communication), a model that can be utilized in an SDS context, or an observation that can inform the design of further experiments/implementations.

(b) The communication feature(s) studied: these features can be prosodic (pitch, loudness, speech rate or vowel duration) , temporal (typically inter-speaker silence duration), lexical and/or syntactic features (usage of same words/syntactic structures by both speakers), gestural/postural (position, gaze, body and head movement) or phonetic (pronunciation).

(c) The speech corpora used: these can be recordings of human dialogues in various contexts, tasks, or settings. A distinction can be made between face-to-face conversations and dialogues without visual contact (e.g. telephone conversations). In addition, there are studies that measure convergence of users towards a synthesized voice in an SDS context. These can either be actual SDS or Wizard-of-Oz implementations, in which an experimenter controls the responses of the system, while the subjects believe that they are interacting with an actual system (see section 2.2.4).

(d) The method of quantification: there are studies that compare speech features of speakers across several dialogues. In this way, a whole dialogue becomes a single data point (e.g the average intensity of a speaker in an entire dialogue). Summary statistics (mean, standard deviation) or regression (between speaker A and speaker B) can be then used to validate the hypothesis of accommodation. In contrast, some studies investigate accommodation *within* a particular dialogue. This approach can be continuous, which usually results in two time series (one for each speaker), or a comparison in "initial" and "final" values of the features which are measured for the first and second half of a dialogue respectively. A third approach is to measure the effect of categorical events (e.g. priming targets – see section 3.5.2) in a sequence of turns or frames after the event, in

order to determine the effect of the prime on the interlocutor's speech.

The studies reviewed in this section are categorized based on (d) above (the method of quantification): Studies based on comparisons of features across dialogues are reviewed in section 4.2. These are further distinguished into studies that measure the average of a feature per speaker for the entire dialogue, and studies that consider specific lexical elements or utterance categories on which the features are measured.

Studies which measure accommodation phenomena within each interaction are reviewed in sections 4.3 - 4.5. Section 4.3 reviews studies in which measurements are based on ratios of successful repetition, a methodology restricted mainly to lexical/syntactic features, as in this case the repetition targets (specific words and syntactic structures) are categorical. A subset of these studies comprises linear regression in order to describe the effect of *distance*, i.e. the frequency or probability of repetition as a function of distance (in seconds or in dialogue turns) from the initial target.

A somewhat unique approach to assessing accommodation phenomena (rhythmic entrainment) within single interactions is reviewed in section 4.4, as it does not lend itself to the categorization followed in this chapter. Time series approaches to describing accommodation phenomena are reviewed in section 4.5.

It is noted that the categorization implied in the outline provided above is not strict, as there is significant overlap across categories (e.g. some studies investigate accommodation phenomena across several dialogues as well as within single dialogues). The categorization only serves presentation purposes. In ambiguous cases, studies have been categorized according to the measurement methodology relevant to the main findings of each study.

Another significant note relates to the definition and theoretical framework disparity that was discussed in the previous section. Since there are various relevant theories, terminologies and lack of universal definitions, the terminology used in each study is also used in its description. The theoretical foundations and the phenomena investigated should become clear from the description itself.

4.2 Comparison of features across dialogues

This section reviews studies that assess accommodation based on comparisons of speech features across several dialogues. One of the methods comprises calculating an average value of a feature per speaker for the entire dialogue and comparing several speaker pairs (section 4.2.1) or the average feature value of the same speaker across several conditions (section 4.2.2). The latter subset

also considers within-dialogue accommodation, by splitting the dialogue in two equal parts (early and late) and comparing feature averages between these two parts. Sections 4.2.3 and 4.2.4 review studies in which measurements are based on specific lexical items or utterance categories, respectively.

4.2.1 Comparison of average inter-speaker pause across dialogues

(Bosch *et al.* 2004a, 2004b; 2005) focused on temporal aspects of turn-taking in human dialogues. The goal of these studies was to investigate the effect of speaker change on temporal features (pause length and overlaps) in corpora with "shallow" annotations, i.e. annotations of speech/silence that contain little or no information about the linguistic content of the utterances. The corpora comprised recorded telephone conversations and face-to-face conversations without any constraints (spontaneous speech). The two features investigated were pause duration and frequency of overlaps. Pauses were distinguished into three types: (a) within-utterance, (b) between utterances within the same speaker turn, and (c) between speaker turns. The annotations were either temporal only, from which turns are defined depending on the temporal organization of speech/silence among the two speakers, or semantic, by incorporation of a basic utterance categorization scheme (propositional vs backchannel with three subcategories each). Silence durations were log-transformed, as this yields a more "bell-shaped" distribution that makes arithmetic means good estimates of the average duration of a speaker for the entire dialogue (see section 8.2.2).

Comparisons of average pause duration per speaker (for each of the three pause categories) in 93 telephone dialogues showed a high correlation for between-turn pauses, as well as a combined set of between-utterance and between-turn pauses. The frequency of overlapping speech at turn exchanges was found to be dependent on sex and dialogue type: a greater proportion of overlap was found for female pairs compared to male pairs, as well as for telephone conversations compared to face-to-face conversations. In addition, turn-exchange latencies were found to have an even two-tailed distribution around a positive peak. The left tail extended into negative values, when duration of overlapping speech is taken into account for turns that initiate before the previous turn of the interlocutor has finished.

(Bosch *et al.* 2005) suggested that the correlation of inter-speaker pause length is evidence of inter-speaker accommodation as proposed in (Pickering and Garrod 2004) and (Giles *et al.* 1992), but also offered the alternative explanation that the correlation might be the result of dialogue or topic liveliness: interlocutors that are engaged in a lively 'chat' are likely to exhibit reduced pause length, thus yielding a medium-sized correlation. One of the main problems identified in (Bosch *et al.*

2005) is that of defining "turns", especially when there is little information on the actual content of the utterance, as is the case in large corpora with shallow annotations.

4.2.2 Comparison of feature averages between first and second half of a dialogue

(Coulston *et al.* 2002) examined amplitude (intensity) convergence of children to simulated educational SDS applications with embodied agents ("animated personas"). The experimental setting comprised a Wizard-of-Oz (Woffit *et al.* 1997) scenario in which children aging 7-11 years interacted with animated characters, while the SDS output was controlled by an experimenter who was in a different location. Thus, children believed the SDS was automated. The TTS voice of the system had two different voice personalities (introvert and extrovert) with different prosodic characteristics (including amplitude). The goals of the study were to (a) examine whether children converge to TTS voices of these, (b) determine whether they do so dynamically, during an interaction, (c) determine whether this happens both in the case of upward or downward movement in order to converge and (d) evaluate the magnitude of convergence.

The children were assigned three tasks. During the first two, the speech of the main character remained constant (introvert or extrovert), while in the third there was a switch in style half-way through the interaction (in one of two directions: from introvert to extrovert and vice versa). There was also contrasting speech style in a sub-character, in order to test *short-term* accommodation. Children engaged in sub-dialogues with the sub-character, which had an introvert voice when the main character had an extrovert voice and vice versa.

Amplitude was measured in voiced regions of utterances, as well as in hand-labeled vowel regions only. A comparison of mean amplitude of children's speech across dialogues showed that they converged to the TTS voice style. The significance was evaluated by a *repeated measures ANOVA*[9] which showed that the children raised their amplitude significantly when interacting with the extrovert TTS voice (higher amplitude) and similarly lowered their amplitude when interacting with the introvert TTS voice (lower amplitude). In the case of the style change, the mean amplitude from the first half of the dialogue was compared to that of the second half and it was found that children showed both upward and downward convergence. Little or no significant evidence of convergence was found for the sub-dialogs with a second character that had a contrasting speech style.

As a measure of the magnitude of convergence, (Coulston *et al.* 2002) used the *percentage increase/decrease in energy,* which can be calculated from intensity. This was calculated for each

9 A variant of the ANOVA method, which calculates a mean and variance from a subset of the population based on a condition; individual observations may satisfy more than one conditions, hence these observations are *repeated*

subject (child) individually, and the increases in the introvert to extrovert condition ranged from 0 to ~300%, with a "grand mean" of 37%. The measurements of amplitude in hand-labeled vocalic regions were slightly more sensitive compared to automatically detected voiced regions, as they generally yielded smaller p-values in significance tests (repeated measures ANOVA).

Following the same methodology, two more studies (Darves and Oviatt 2002; Oviatt *et al.* 2004) extended the set of prosodic features, including durational and temporal features (utterance duration, number of within-utterance pauses, speech rate and response latency). The highest magnitudes of accommodation (as percentage increase/decrease of energy, log transformed duration, or speech rate) was found for within-utterance pauses (both in number and duration). In general, children were found to adapt all of the aforementioned features, depending on the style of the TTS voice (introvert vs extrovert). Less concrete evidence of convergence was found for the sub-dialogues with a sub-character with contrasting TTS voice, except for the intra-sentence pause patterns. In addition, little or no evidence was found of any effects of age, gender or personality match between child and TTS voice (introvert or extrovert).

(Coulston *et al.* 2002; Darves and Oviatt 2002) and (Oviatt *et al.* 2004) proposed that these findings can be helpful in SDS design in order to guide children's speech to prosodic behaviour that is easier for ASR to handle (e.g. low amplitude in children speech is a problem in speech recognition).

A series of studies (Suzuki and Katagiri 2003, 2004, 2005) examined prosodic alignment/synchrony of users' features (intensity and response latency) with the respective features of an SDS voice (pre-recorded prompts). The goal of the study was to compare the findings with those of previous studies on human-human dialogues, in order to find evidence of similarities or difference between these two settings. Recordings of adult Japanese speakers interacting with the SDS in a Q&A quiz scenario were used for the analysis. Since the dialogues were "half-duplex" (because of the Q&A structure of the dialogues), turns were annotated in a straightforward way and the response latency between subject and speaker turns was measured. The average intensity of the user utterances was also measured. During the first half of a quiz, participants interacted with the original SDS voice, while in the second half one of the the prosodic features was modified (±3dB in intensity or ±30% in response latency) or both were left constant. The mean intensity and response latency were calculated for each speaker and half-dialogue.

In the follow-up statistical analysis, all three studies used t-tests to find whether the prosodic features studied changed significantly in the increasing and decreasing conditions. The results showed alignment in both directions (increasing and decreasing) and no significant change in the constant conditions. However, it was found that the changes were statistically significant only in the

increasing condition for intensity and the decreasing condition in response latency, i.e. only in one direction for each a/p feature.

These results partly evaluate the predictions from human-human dialogues, where alignment of both features occurs in both directions (Jaffe and Feldstein 1970), as was also found for human-computer interaction in other studies (Coulston *et al.* 2002; Darves and Oviatt 2002; Oviatt *et al.* 2004) that followed a very similar methodology for measuring alignment of the same features. The comment of (Suzuki and Katagiri 2005) on this difference was that the latter studies used Wizard-of-Oz scenarios rather than actual SDS for their tests. However, there were other differences: the former studies (Coulston *et al.* 2002; Darves and Oviatt 2002; Oviatt and Seneff 2004) used synthesized TTS voices rather than pre-recorded prompts, and children rather than adult subjects. In addition, the type of interaction was different in the two cases: a talking agent educational environment in one case and a Q&A test in the other.

(Suzuki and Katagiri 2005) concluded that alignment of users to SDS is a global phenomenon that can be utilized to serve SDS efficiency, in relation to ASR component: a system can adapt its amplitude in order to make the user voice converge towards a value that yields better performance of automatic speech recognition.

4.2.3 Comparison of features measured on specific lexical elements

(Pardo 2006) conducted a study on phonetic convergence. Several theoretical foundations are discussed, such as priming, entrainment or influence of social factors. The goal of the study was exploratory: to provide evidence supporting/rejecting the several hypotheses. The speech material was recorded during a task-based experiment (map-task), during which one of the subject had to draw a path on a map that contained landmarks, based on the instructions of the other subject (who had a complete path). The efficiency of task execution was assessed by superimposing the two paths on a 1cmX1cm square grid and calculating the number of the squares that the two paths had in common. The phonetic similarity between the two speakers was assessed on identical lexical elements (names of landmarks on the map) by perceptual listening tests, in which (different) subjects were asked to make a forced choice of similarity to a sample utterance, based on *pronunciation* (as this tended to draw the focus of listeners on the phonetic content, rather than prosodic or voice quality or any other features). This was done for utterances recorded before, during, and after the task.

The ANOVA method was used to assess convergence based on number of factors such as (a) talker role (information giver vs information receiver), (b) sex (male vs female), (c) persistence (pre-task

vs during-task vs post-task), and (d) timing (first half vs second half of the dialogue). Conversational partners were found to converge phonetically during the task (compared to pre-task) and more so over time (early vs late in the dialogue). In addition, convergence persisted beyond the task (post-task instances were judged more similar than pre-task instances). Further, information givers were found to converge more towards receivers than vice versa and male pairs showed more convergence than female pairs. (Pardo 2006) presents a detailed discussion related to theories of episodic memory, perception-production link, entrainment and social factors, in view of the experimental results: for entrainment, the degree of "coupling" plays an important role, thus it is suggested that convergence is less likely in relaxed interactions than in task-based scenarios with increased cognitive load; relative coordination, as in the Interactive Alignment model, is a more plausible explanation of the phenomena; the link between perception and production is not automatic; and situational constraints impose restrictions on convergence in relation to social factors.

4.2.4 Comparison of features measured on specific utterance categories

(Ward and Nakagawa 2004) explored speech rate adaptation in human conversations and proposed a methodology for IVR implementation. The corpus in the study consisted of 508 recorded telephone directory service dialogues, in which there was information provided to the user in the form of a series of digits (telephone numbers). This was not an actual service, but an experimental set-up in which human agents (with prior customer service experience) were used. It was hypothesized that the (human) agents delivered the digits faster or slower depending on (a) the users' initial speech rates, and (b) the user's response latency after the initial greeting of the agent, which is seen as a measure of hesitation. These hypotheses were tested on a subset of dialogues that were (a) previously rated "good" by the subjects (callers), and (b) the digit delivery pattern was the most commonly occurring (a confirmation after each group of digits). Speech rate was measured in morae/second. A mora was defined as "roughly a syllable" in (Ward and Nakagawa 2004): an approximation of two morae per double vowel, and one mora per single vowel, syllabic nasal or geminate consonant was used. This resulted in user speech rates ranging from 6 to 10 morae/second.

Significant correlations were found between among both user speech/rate and response latency to the agent's initial greeting on one hand, and the duration of delivery from the agent on the other. A linear model, with both factors as independent variables and the agent delivery duration as a dependent variable, was calculated by multiple regression (least squares). This model was then

tested in the design of an IVR implementation, in which the conversation was handled by a human agent up to the point of the information delivery (the actual digits). The novelty of the system was that the user speech/rate and response latency were measured on-line, so that the final delivery (automated) could be implemented based on the previously fitted linear model. The evaluation showed a significant correlation between the predicted duration for the system and the actual duration in the corpus.

(Ward and Nakagawa 2004) noted that they did not evaluate the system on-line with real users, considering two issues. First, that the system needs a "sanity check", in order to avoid producing too long or too short deliveries (based on erroneous parameter measurement online). Second, users do not tend to confirm groups of digits when the delivery is performed by a machine. The conclusion was that speech rate adaptation should find numerous applications in SDS.

(Bell *et al.* 2003) examined user prosodic behaviour during interaction with a Wizard-of-Oz implementation of an SDS. The goal of the study was to investigate users' adaptation of their speech rate during mis-recognition and other errors, in order to explore possible design strategies for SDS. A common problem in real SDS environments, is that users typically hyper-articulate their speech after a speech recognition error, since that strategy "works" in human-human dialogues. Unfortunately, the same strategy has the opposite effect in SDS, as ASR performs badly on hyper-articulated speech (Bell *et al.* 2003).

The study utilized a Wizard-of-Oz scenario (Woffit *et al.* 1997), in which subjects (members of the general public) interacted with either a fast or slow version of the SDS. The goal of the task was to aid an animated character complete a sorting task comprising geometric shapes of different colors. During the interaction, experimenters deliberately introduced errors, such as mis-recognitions. The subjects either repeated or rephrased their utterance. The fast and slow version of the system were implemented by modifying the speech rate of the original pre-recorded prompts by ± 30%. The measurement of user speech rate was segment duration, which was calculated using an automatic alignment algorithm, the output of which comprised an annotation of words and phonemes. Stressed syllables were also annotated. A z-score technique normalized the durations for inherent duration and effects of stress.

ANOVA tests were used to determine the effects of user turn, system speech rate and lexical content on the user speech rate. User turns were distinguished into original user utterances, re-phrasings, and repetitions. Lexical elements were distinguished into descriptions of the shapes (color, shape, position) and speech rate was either fast or slow. All three independent variables were found to have a significant effect on user speech rate, with words describing color being the only lexical content

that was pronounced significantly slower. It was also found that user turn type (original, rephrasing, repeating) also had a significant effect on within-utterance pause length in user speech. Finally, the speakers spoke slower to the slow version of the system, thus validating the hypothesis of convergence to the TTS voice.

(Bell *et al.* 2003) observed that users adapt their speech rate unknowingly according to the speech rate of the system, but also depending on the dialogue context. In order to handle system errors they slow down but, as soon as the system recovers from the error, the users "speed up" quickly and the dialogue flows smoothly.

4.3 Measurements of successful repetition

This section reviews studies that measure successful repetition of targets, which are typically lexical or syntactic elements, although prosodic targets have also been defined in some of the studies (Ward and Litman 2007b, 2007a). Section 4.3.1 reviews two studies that have measured successful repetition ratios of lexical elements. Section 4.3.2 reviews two studies which, in addition to successful repetition, have also used linear regression in order to measure the effect of distance on the probability or frequency of repetition.

4.3.1 Successful repetition ratio

(Brennan 1996) studied lexical entrainment in recordings of spontaneous speech, as well as adoption of system terms (lexical convergence) by users of speech interfaces. The goal of the study was to investigate differences and similarities between these two processes, and implications of this for SDS, especially in relation to the vocabulary problem: the wealth of language is a problem for SDS, because a user may adopt several terms to describe the same concept. Lexical entrainment (or convergence) is a possible way of encouraging the user to use specific terms (by presenting this vocabulary to the user), thus shortening the list of candidate words that the ASR and ALU components have to process, which in turn would result in increased efficiency.

The spontaneous speech recordings were acquired using a task experiment, which involved two participants who could converse without visual contact and had to line up identical sets of picture cards in the same order. The purpose of these experiments was to further investigate previous theoretical predictions on lexical entrainment (Clark and Wilkes-Gibbs 1986; Brennan and Clark 1996). The latter explain lexical entrainment through "conceptual pacts", or implicit "agreements" between interlocutors on terms that describe concepts in the discourse.

(Brennan 1996) conducted a series of wizard-of-oz experiments (database query), in which a (simulated) system employed two different correction strategies in order to encourage the user to adopt its terminology: "embedded" and "exposed". Embedded corrections are repetitions of the query by the system with substitution of the user term with the system term, while exposed corrections are explicit clarification requests of the system that contain the system term (e.g. "did you mean */term/ ?)*. A speech-based, as well as a text-based interface were used.

The measurements comprised a ratio of successful adoption of the system term by users over the total amount of user turns. Also, the effect of *delay* (whether the user response came immediately after a correction or after several utterances) was investigated. In both text and speech cases, there was significantly more lexical convergence for the immediate condition (compared to delayed) and exposed corrections (compared to embedded). However, convergence was significant in all cases. (Brennan 1996) suggested that convergence only in the immediate condition would imply autonomous entrainment, while frequent convergence in the delayed condition would imply a more strategic process. The study concluded that (a) a system should output only terms that it can process as input, (b) should be consistent in its output and documentation, (c) repairs are essential, as shown from the higher convergence to exposed corrections, and (d) a system could adopt terms proposed by the user by adopting grounding strategies.

In (Fais 1996), lexical accommodation was investigated in view of human-machine interface design. Three experimental scenarios were conducted, in which a conversation was either (a) direct monolingual, between English-speaking subjects and conference "agents", (b) bilingual, between English-speaking and Japanese-speaking agents, mediated by a human interpreter, and (c) mediated by a simulated machine translation system. The goal of the study was to study lexical accommodation in these three contexts in order to determine the effect of (1) desire for social approval, and (2) difficulty of communication, on the degree of accommodation. The measurement was a ratio of the number of (different) words spoken by both speakers over the overall number of (different) words in each dialogue. The direction of accommodation was assessed by defining that a speaker who uses a word previously spoken by the interlocutor is the one who accommodates.

The results showed significant accommodation in all three scenarios. The highest accommodation occurred in the human-mediated scenario where, according to (Fais 1996), both social factors and communication efficiency are important. Higher accommodation was also found for the machine-mediated scenario, when compared to the direct dialogue scenario, despite the fact that social factors were irrelevant. In addition, accommodation was equal between interlocutors in the direct dialogue scenario, but in the other two the client accommodated to the agent. (Fais 1996) attributed

this finding to the fact that clients perceived interpreters (either human or machine) as having the dominant role. Thus clients accommodated to the lexical choices of the interpreters, in order to improve communication efficiency. The latter conclusion is in agreement with (Brennan 1996). (Fais 1996) also suggested that higher accommodation in SDS can be encouraged by use of an animated face or "persona", replicating the human-interpreted setting (that shows the highest accommodation).

4.3.2 Linear regression of repetition over distance

(Reitter *et al.* 2006) explored priming of syntactic structures, in order to test various predictions of the Interactive Alignment Model (Pickering and Garrod 2004) that was outlined in section 3.5. Spontaneous (telephone) speech and task-oriented speech (a map-task, which was described in section 4.2.3) were used in order to test the effect of the situational constraints of the task on the degree of priming. The syntactic trees of all utterances in the corpus were converted to phrase rules and an algorithm search for repetition of these rules was conducted. Any sentence could be a valid candidate for a prime or target for priming (repetitions of entire phrases were excluded). *Distance* (expressed in number of turns or seconds) of priming was also taken into account. In addition, a distinction was made between *comprehension-production* (CP) priming, where one speaker produces the prime and the partner produces the target, and *production-production* (PP) priming, where both prime and target are produced by the same speaker.

Statistical analysis is performed by use of generalized linear mixed effects regression models (GLMM). This regression approach allows the calculation of coefficients of linear models, such as a model of the probability of repetition of a prime, based on discrete factors (such as type of corpus) or continuous explanatory variables (such as distance). The maximum distance used was 25 turns or 15 seconds. There were various outcomes from this study. The probability of priming was found to decay with distance in both corpora, and significant PP priming was found in both corpora. In the case of CP priming, higher confidence was found for the map-task corpus, when compared to spontaneous speech. This, according to (Reitter *et al.* 2006), validates the hypothesis of the Interactive Alignment Model that syntactic priming leads to semantic priming: when the cognitive workload is increased (task corpus), speakers reproduce each other's syntactic structures in order to align their situational models with less effort. In the unconstrained spontaneous speech, the cognitive workload is less, thus the speakers are less eager to adopt their partners' syntactic structures, but they still do so to a lesser extent mechanistically.

Following (Reitter *et al.* 2006), (Ward and Litman 2007a, 2007b) investigated dialog convergence

in relation to learning in tutorial sessions with a human tutor and an intelligent tutoring system (SDS). The features studied were lexical (word repetition) and prosodic (F0 and Intensity). The theoretical background of the studies was based on the Interactive Alignment Model (Pickering and Garrod 2004).

The measure of lexical convergence was the count of different word tokens repeated by the student in a window of up to 20 turns after the tutor's utterance (prime). For prosodic features, the minimum, maximum and average F0 and Intensity of the tutor's utterance were considered primes if their z-score normalized values where greater than one (an arbitrary threshold of one standard deviation). Again, the response of the student to the prosodic prime for a window of up to 30 turns (to capture variation in intensity) was measured. The effect of distance on the number of repetitions (either lexical or prosodic) of a prime in the speech of the tutor was measured as the slope of a line fitted by linear regression (least squares). The slopes typically have a negative value, which is an indication that prime repetition decays over time. The significance of this was assessed by calculating a p-value, as an indication of the probability of fitting that line if there was no effect of distance.

In order to assess the effect of convergence on learning, (Ward and Litman 2007a) used a corpus of students who completed two physics tests, one before and one after a tutoring session with a (human) tutor. Thus, the learning outcomes from the tutoring session were quantified by means of test-scores. An automatic feature selection algorithm (stepwise regression) was used to find which features, if any, affect the learning outcomes. The only factors that were identified by the algorithm were lexical repetition and response on mean intensity primes, for a window of 20 turns. The identified models were then tested on a different corpus, which contained dialogues between students and an automatic intelligent tutoring system. The latter was following the same tutoring session layout and procedure as in the sessions with a human tutor. The models remained significant in the test data, although there were some unexplained differences (e.g. change of sign in some coefficients). (Ward and Litman 2007b) concluded that there is evidence of a relationship between convergence and learning, despite contrasting differences of the models in the two corpora, which can possibly be explained by the differences in speech style between human-human and human-computer conversation.

4.4 Assessment of latency distribution

(Benus 2009) studied rhythmic entrainment of syllable and pitch accent timing in human dialogues and, in particular, the predictions of the “coupled oscillator model” (Wilson and Wilson 2005),

which are (a) isochrony in turn-internal chunks, (b) entrainment across turn exchanges, (c) latency distribution should be bi-modal with two peaks around zero (when considering overlap turn exchanges as negative latencies) and a valley at zero, and (d) that the entrainment should persist without signal transmission (when both speakers are silent) for a period up to (roughly) one second, after which there should be more simultaneous starts observed. A corpus (American English) of young adults recorded while playing games in separate isolation sound-proof booths, and communicating via audio channel only, was used in this study.

Turns were categorized using the temporal scheme of (Beattie 1982) in order to determine turn types. This scheme considers seven categories of "speech chunks": (1) backchannel, (2) backchannel with overlap, (3) smooth switch, (4) overlap switch, (5) "butting-in" (or unsuccessful interruption), (6) interruption-by-pause, and (7) Interruption by overlap. In addition, (Benus 2009) defined two additional labels, namely (8) continuation of the same speaker after a back-channel, and (9) simultaneous start.

(Benus 2009) used syllables and pitch accents as the rhythmic units of speech, following the proposal of (Wilson and Wilson 2005). The utterances were transcribed using the TOBI scheme (Silverman *et al.* 1992) and the time of maximum energy was used as an estimate of the pitch accent location in accented syllables. Correlations among syllable durations or pitch accent latencies were used to test the several hypothesis (a-d above) of (Wilson and Wilson 2005). In addition, a "phasing measure" was defined as latency/chunk-rate to test the particular hypothesis that latency before initiation of vocalization depends on the rhythm of the preceding speech chuck. The latencies of the 9 categories defined above were plotted as histograms, on top of which the phasing measure was plotted as a smooth curve. In this way, the hypotheses of the model of (Wilson and Wilson 2005) could be validated by inspecting the histograms of latencies between chunks, based on syllable boundaries or pitch accent locations, if peaks could be found at specific latencies.

The results showed weak support for the model of (Wilson and Wilson 2005) especially in relation to hypotheses (b) persistence of rhythmic entrainment across turn exchanges, and (c) bi-modal distributions of response latencies. (Benus 2009) attributed this to several possible factors, such as the type of corpus (task-based or spontaneous) or timing units used (syllable and pitch accent may not be the most suitable) and concluded that perhaps the key assumptions of the model ought to be rethought. It is noted that hypothesis (c) was also not validated in (Bosch *et al.* 2004b) where one positive two-tailed peak was found with the left tail extending into the negative values. Further, the results of (Benus 2009) are in agreement with those of (Bosch *et al.* 2004b, 2004a; 2005) regarding accommodation of pause latency at turn exchanges.

4.5 Time series measurements of accommodation

This section reviews studies that have utilized time series analysis in order to describe inter-speaker accommodation. The approaches mostly vary in two ways: the method of obtaining the points, and the statistical analysis performed on the time series. Points can be obtained either by averaging a feature over the duration of an utterance, or by measuring the feature at specific elements (words, syllables) or utterance categories (e.g. only at the beginning of turns). In case of continuous phenomena, such as head movement, points can be obtained by direct sampling (without averaging). The statistical analysis can also vary from making inferences simply from observing simultaneously plotted time series of two speakers to more sophisticated statistical methods, such as as cross-correlation analysis, lag regression analysis, recurrence analysis and spectral analysis.

4.5.1 Time series plots of utterance-based feature averages

(McRoberts and Best 1997) studied the prosodic convergence of an infant to her mother and father at various ages (3-18 months) in the context of validating hypotheses from CAT (Giles *et al.* 1987; 1992). In particular, F0 of infant and parent were measured for interactions recorded at home on a weekly basis. The mean F0 was calculated for each utterance and a grand mean for the entire interaction (15-20 minutes) was calculated from these. A weighted mean for each utterance was also calculated by multiplying each utterance by its duration and dividing the sum of cross-products by the grand mean of the dialogue. This was done to exclude the possibility of bias introduced by correlation of utterance duration and F0. Using ANOVA, the authors found that the parents raised their F0 when interacting with the infant, and there was an effect of the infant's age on their F0, although different for each partner (a more “linear” adaptation was found for the mother). The infant's F0 was not found to change when interacting with either parent, compared to when she was alone.

Further, the (McRoberts and Best 1997) examined F0 convergence within single interactions, by plotting the mean F0 of infant and parent as two time series. The “time” variable was the utterance number. As the utterances were of various lengths, it was not possible to determine time intervals from the plots in (McRoberts and Best 1997), nor are the points of the two series *synchronous*. The study discussed the apparent synchronous movements in F0 that can be observed, although it pointed out the difficulty to infer conclusions from the data. (McRoberts and Best 1997) also noted that the scale of the Y-axis affects the apparent “similarity”: increasing the Y-axis “resolution” by a factor of 2 revealed that previously “similar” points were actually very distant. The results were interpreted as an indication of prosodic (F0) convergence, as defined in (Giles *et al.* 1992).

4.5.2 Simultaneous time series plots of activity in multiple modalities

(Campbell 2009) explored multimodal synchrony in video-captured conversations of more than two interactants at a time. The goal of the study was to illustrate the process of *active listening* and *synchrony*, by presenting evidence of simultaneous activation of interactants across features (or modalities) of speech, gesture, and posture, in a continuous representation (in contrast to half-duplex, or "ping-pong" representations). In particular, vocalization, hand gestures, head movement and body pose were the features studied. The corpora comprised telephone dialogues in Japanese, as well as video recordings acquired with a 360^0 camera position at the center of a table around which four or five participants were sat. The head and movement position was captured dynamically from the video recording and the movements were automatically tracked by a 2-D algorithm tracking lateral and up-down head movement with additional correction based on size for the third dimension (back-front). The vocalization intervals of each speaker were manually annotated. The annotation and analysis of gestures employed the MUMMIN coding scheme (Allwood *et al.* 2007).

Based on visual assessment of chronographic representations of the telephone dialogue recordings, (Campbell 2009) pointed out that (a) periods where the dialogue is dominated by either speaker are likely to be rich in propositional content, while short bursts of overlapping speech are characterized by backchannels expressing understanding or agreement and other such functional gestures of feedback, and (b) that it is very difficult (if not impossible) to define "turns" or "turn-holders" in these cases of short overlapping segments that frequently occur in natural conversational speech. (Campbell 2009) posited that dialogue is a synchronous interaction in which *both* participants continuously participate by a process of *active listening*. The data from the video recordings verified this hypothesis, as high correlation was found in action, gaze and pose among four interactants. In addition, high correlation between the imagery analysis data (which is automatic) and the chronographic representation (manually segmented) of these dialogues was also found, as bursts of movement and overlapping speech were identified as points of high activity in the interaction. These "activity peaks" were found to be common to all interactants most of the time, which was proposed as evidence of synchrony in the interaction.

(Campbell 2009) concluded that these findings provide evidence that interactants participated positively in the dialogues, and that their multimodal synchronization was a result of this. In addition, the automatic feature extraction from the visual data can be very useful in detecting "activity peaks" without the need for manual annotation. In an SDS context, these peaks in activity can be indicative of topic changes or any other significant events in the discourse, therefore SDS could employ automatic activity peak detection in order to be aware of important discourse events.

4.5.3 Time series plots of features measured on specific targets

(Kakita 1996) studied F0 convergence during dialogues between humans, in order to test the validity of theoretical predictions and based on previous work of the same author on other prosodic features (speech rate and pause duration). The speech material consisted of 18 scripted question-answer pairs, which were designed so that a specific vowel was always pronounced in the same, non-emphatic position, in all questions. After each answer, the subjects exchanged roles, so that points were acquired for both (25 utterances each). The subjects were male adult Japanese students.

The F0 was measured on the target syllable, and the points were plotted as two time series, with the X axis representing question-answer pairs, numbered 1-50, and the Y axis representing F0. Trend lines were fitted to each series by linear regression (least squares). By visual observation of the slopes of the fitted lines, (Kakita 1996) distinguished three patterns: (a) convergence, (b) divergence, and (c) parallel movement (unaffected). Further investigation showed that when speakers had a small difference in their initial F0, they tended to diverge. In contrast, large initial difference lead to more cases of convergence.

Based on these (briefly summarized) findings, (Kakita 1996) hypothesized various possible causes of convergence as a function of initial difference, and identified a region of 5-20Hz of initial difference that is possibly optimal for inducing convergence, although that could not be verified due to the small amount of data in the study. Across dialogue comparisons of initial F0 for subjects that took part in more than one dialogue showed little evidence of per-partner adaptation of F0.

4.5.4 Calculation of lag-zero coefficient

(Nishimura *et al.* 2008) studied the relationship between synchronous prosodic variation, or "synchrony tendency" and perceived "liveliness", "familiarity" and "frankness", in a corpus of spontaneous dialogs in Japanese. The goal of the study was to find useful features that can be used in making an SDS voice more pleasant. The prosodic features studied were F0, F0 range, intensity, intensity range, speech rate, speech rate range. Averages of these features were calculated for each utterance and these were plotted as a time series, with each point located at the center of each utterance in time. Contemporaneous points were obtained by means of linear interpolation for one of the speakers (chosen randomly) at the times of the points of the other speaker.

Significant positive lag zero correlations between the two speakers were found for all features in 389 out of 508 1-minute fragments taken from 7 dialogues. (Nishimura *et al.* 2008) suggested that this is evidence of high synchrony tendency. The results of the time series analysis were combined

with questionnaire results from a perceptual study, in which listeners were asked to rate the same corpus for “liveliness”, “familiarity” and various other descriptors. Again, high correlation was found between higher ratings and higher synchrony tendency, especially for F0. In addition, the correlations were increasingly stronger and found in a greater percentage of the dialogue extracts when the rating for “liveliness” was higher.

Since SDS voices that sound familiar and lively are desirable, the authors (Nishimura *et al.* 2008) proposed a multivariate regressive model that monitors the user's prosodic feature (pitch, intensity, speech rate) and adapts the same feature on the voice of an SDS, in order to match the behaviour observed in human dialogues. The parameters of the model are the average value of a prosodic feature for the last N turns and a time constant K that specifies the delay of the system (how quickly it adapts). High correlation of the model values calculated from speaker A with actual values of speaker B were found. However the 'optimal' parameters computed did not allow any delay for the system to converge to the user in some cases. (Nishimura *et al.* 2008) concluded that the model follows the user passively, but it should actively change its prosodic behaviour depending on the context.

4.5.5 Pearson coefficient

(Edlund *et al.* 2009) examined convergence and synchrony of pause (between utterances of the same speaker) and gap (pause between utterances of different speakers) length across two speakers in six spontaneous dialogues. The study was proposed as a proof-of-concept for the proposed methodology, which is a time series approach to measuring convergence continuously. The overall goal stated is to produce a model that can capture the dynamics of convergence on-line and in real-time (in view of implementing similar behaviour in SDS). The speakers were recorded in free face-to-face conversation. The audio channels were processed with a VAD algorithm that made the speech-silence decision. The resulting durations (in milliseconds) of the gaps and pauses were then transformed into the log domain, which is based on previous findings (Jaffe and Feldstein 1970) that the distribution of silent interval lengths is positively skewed, thus making arithmetic means overestimates.

(Edlund *et al.* 2009) distinguished between *convergence/divergence* and *synchrony.* The first was defined as the decrease/increase of the difference in duration of pauses or gaps across two speakers over time; in other words, speakers were considered to converge when the similarity in gap and pause duration *increased over time.* The second was defined as contemporaneously similar variation of pause or gap duration across the two speakers, i.e. whether the speakers' variations in pause and

gap duration show similar local trends. (Edlund *et al.* 2009) used a 20-point feature window (the average of the last 20 pauses or gaps) with varying length: some windows had a length of less than 1 minute or more than 6 minutes, with most frequently occurring lengths of 2 minutes for pauses and 3.5 minutes for gaps. *Linear interpolation* was used in order to compare values from one speaker to interpolated values of the other speaker at that exact time, randomly chosen each time. (Edlund *et al.* 2009) reported that exchanging the speaker data (interpolating speaker B instead of speaker A values) had a negligible effect on the results. Statistical evaluation of synchrony was performed using the Pearson correlation coefficient between the values of two speaker series for each dialogue. For convergence/divergence, the *differences* between the values of speaker A and the interpolated values of speaker B were correlated (Pearson coefficient) with the time of their occurrence in each dialogue.

Significant correlation was found for both tests and for both gaps and pauses in a portion of the dialogues. However, few dialogues from the overall set of 6 showed significant convergence and even fewer showed divergence. (Edlund *et al.* 2009) noted that the hypothesis of convergence (or accommodation of pause/gap length) being a global phenomenon was not validated. Synchrony was more evident according to the same results, as most dialogues showed strong positive correlations. Some (weak) negative correlations were also found. The conclusion of (Edlund *et al.* 2009) was that possibly there are other factors of variation in pause and gap duration, which "override the general synchrony of the exchange".

4.5.6 Lag regression analysis

(Jaffe *et al.* 2001) studied rhythmic "coordination" in mother-infant communication. The goals of this study were to (a) describe the vocal rhythms in such interactions based on previous work on speech rhythm (Jaffe and Feldstein 1970), (b) describe coordination of vocal rhythms in these interactions in terms of their significance and bi-directionality, (c) predict infant development (attachment and cognition) at age 12 months from coordination at age 4 months, and (d) explore whether familiarity of partner or environment has any effect on coordination (using stranger-infant, mother-stranger and home-lab control conditions). The theoretical bases are that rhythm is inherent in speech and interaction (a mechanistic/autonomous approach) and previous studies on the effect of mother-infant coordination on infant development.

(Jaffe *et al.* 2001)used a transformation of the speech signal into an on/off (binary) series of points sampled every 250 milliseconds, which is the smallest time unit in the analysis. Thus the only information in the series is whether either speaker (mother, infant or stranger) is vocalizing or not

(on or off) during the 250 millisecond time-frames, which are represented as points in time series. (Jaffe *et al.* 2001) used two rhythmic features: the beat, as the sum (V+P) of average vocalization (V) and average pause duration (P) in a time frame or turn, and the (V/P) ratio. The latter is proposed as a measure of extroversion/introversion due to the fact that extrovert speech is more "lively" and is thus characterized by shorter pauses in general, in contrast to introvert speech that is more hesitant. A turn of speaker A is defined as beginning when A starts vocalizing alone and ends when B starts vocalizing alone. Five vocal states are defined: (1) continuous vocalization, (2) pause, (3) switching pause (at turn exchanges), (4) non-interrupting overlap and (5) interrupting overlap. The last two categories are distinguished by considering whether the initial turn holder (before the overlap) retains or gives up the turn respectively. The switching pause is defined as belonging to the speaker whose turn it *terminates*. (Jaffe *et al.* 2001) distinguished between (and separately analyzed) *rhythmic entrainment* beats and (V/P ratios), as well as *coordination* (of average durations of vocal states) across speakers within each interaction.

The statistical analysis comprised lag regressions (excluding lag zero) between time series of average duration of vocal states (frame length 5 seconds) and turns (frame length 30 seconds). across speakers in each dialogue. Exchanging data from each speaker as dependent/independent variables, (Jaffe *et al.* 2001) calculated a coefficient of *coordinated interpersonal timing* (CIT) index, as the strength of regression R^2 between the series of each speaker and the *lagged* series of the other speaker. A series of 12 lags (accounting for a period of one minute) were considered in order to assess CIT. (Jaffe *et al.* 2001) considered each speaker's CIT in order to assess whether coordination was uni-directional, bi-directional, or absent, in case one, both, or none of the CIT indices were found significant. Coordination was considered present if the CIT for at least *one* of the vocal states was significant. Further statistical analyses (MANOVA[10] and multiple linear regression) were performed in order to test the effect of setting (home or laboratory) or novelty of partner (mother-infant vs stranger-infant), as well as to infer whether coordination between mother and infant at age 4-month has any effect on infant development. The latter was assessed by a series of specialized observation tests (Jaffe *et al.* 2001).

The general results were (a) non-significant entrainment in beat cycles and V/P ratios, except for familiar partners (mother-infant) at familiar settings (home), (b) significant coordination in the largest percentage of cases, the magnitude of which could be used to predict development outcomes, (c) increased bi-directionality in adult-infant interactions (compared to adult-adult), (d) increased bi-directionality when novelty (of partner and site) is introduced, (e) positive correlation of switch pause and overlap across speakers, (f) negative correlation of pause and vocalization. The

10 Multivariate ANOVA

latter two findings were attributed to convergence and complementarity in temporal (rhythmic) features. Finally, (Jaffe *et al.* 2001) estimated an *optimal lag* which accounted for the largest amount of variation (R^2) in the calculation of CIT from the 12 five second lags. This was found to be in the region 20-30 seconds (lags 4-6) in most cases. This was proposed as a recurrent rhythmic cycle that is inherent in mother-infant interaction.

4.5.7 Spectral analysis of filtered series

(Buder and Eriksson 1997; 1999) studied synchrony of F0 and Intensity in human dialogues. The goals of the studies were to investigate whether synchrony is a persistent, universal or language-specific phenomenon and its relationship to transitions from each speaker to his/her partner.

The corpus consisted of four dialogues, two in American English and two in Swedish, with a male pair and a female pair in each language. One approximately half-minute extract from each of these recorded dialogues was analyzed in order to measure F0 and Intensity synchrony. The prosodic data was extracted at a rate of 240 times per second and further down-sampled by 3-point and 5-point smoothing. Median and mean smoothing was used for F0 and Intensity, respectively. (Buder and Eriksson 1997) pointed out that this process was required in order to exclude micro-prosodic variations, recording artefacts and algorithmic failures. The result of the process was a number of time series comprising 16 samples of F0 and Intensity per second. These were organized into 128-point (8 second) frames with an overlap of 48 points (3 seconds). At each point, the F0 and intensity were normalized to the frame average (and overall sample average) and missing points for the F0 (in non-voiced regions) were zero padded. Spectral analysis (FFT[11] of the filtered signal) revealed periodic patterns in the variations of both F0 and Intensity, to which the authors fitted sinusoidal models.

By observation of the plotted models, superimposed on the prosodic data, (Buder and Eriksson 1997) found that the periodic pattern of one speaker, who dominated the conversation for a part of the dialogue sample, persisted (with aligned period and phase) across the turn exchange to the speech of the second speaker. In signal processing terms, the sinusoidal model fitted to the prosodic data of the speaker who released the floor, fitted well with the prosodic data of the second speaker who took the floor. This behaviour was observed in all four dialogue samples in the study. (Buder and Eriksson 1997) reported that the most typical cycle (period) for the fitted models was 4 seconds for Intensity and 2.5 seconds for F0. (Buder and Eriksson 1997; 1999) concluded that these findings are an indication that rhythmic alignment (or synchrony) in dialogues may well be a universal,

11 The Fast Fourier Transform (FFT) is a computer algorithm that is used to calculate the Discrete Fourier Transform of a signal. The result is a transformation of a signal to the *Frequency Domain.* See: (Rabiner and Schafer 1978)

language-independent phenomenon and proposed further work, in view of natural interaction in SDS applications.

4.5.8 Recurrence analysis

(Richardson *et al.* 2008) studied postural and gaze features, and reviewed a body of previous work in this area. The general goal was to investigate entrainment of conversants' body swing and eye focus, when standing in upright position. In a series of experiments, subjects were involved in several tasks which were designed to produce spontaneous speech, such as watching sitcoms and discussing their favorite characters, discussing a painting, or performing tasks in a common area through wall-mounted monitors, with or without visual contact with other subjects. During these experiments, the body swing (lateral movement of upper body in upright position) and eye movement and focus were recorded continuously.

Statistical analysis of the resulting time series was performed by means of recurrence analysis, a method which, according to (Richardson *et al.* 2008), is more straightforward in revealing recurrent (or cyclic) patterns by observation of *recurrence* plots. A point is registered on a recurrence plot only when events that occur at fixed intervals (recurrently) are sufficiently "similar" (within a preset threshold). Thus, the density of points registered along lines that represent specific periods yields the amount of recurrence for that period. The density can be expressed as a *percent recurrence,* the proportion of points registered on the plot vs all possible points. An extension of this method to bi-variate time series (which comprised cross-recurrence plots and percent cross-recurrence measures) was used to assess coordination among behavioral patterns of two participants.

Interestingly, (Richardson *et al.* 2008) found coordination of body swing even when the subjects were facing away from each other (interacting through monitors on opposite-facing walls), or when there was no visual contact (subjects interacting through monitors without visual contact). In addition, eye movement (gaze) coordination was found not only between partners in an on-going conversation, but also between listeners and speakers when the former were listening to a recorded description of a painting. (Richardson *et al.* 2008) concluded that there is transmission of rhythm through speech, and that this is not only a by-product of interaction but also has an effect on its outcome. (Richardson *et al.* 2008) proposed some evidence that common ground (Clark and Schaefer 1989) is relevant to coordination of gaze, as listeners could answer questions about the painting correctly more often when their gaze was coordinated to that of the speaker.

4.6 Discussion

This chapter has reviewed various methods of measuring accommodation phenomena in various modalities. Regardless of the theoretical foundations or goals, each study measured accommodation in one or more verbal or non-verbal features (see Table 4.1). It was mentioned in section 4.1 that these can be broadly categorized into across-dialogues and within-dialogue measurements.

Across dialogue measurements are the most robust method, as the whole dialogue is used to calculate an average value of a feature: if the dialogue is long enough, then the arithmetic mean can be safely assumed to be unbiased by some event that occurred during the interaction causing unusual behaviour which deviates from the mean. Provided that a sufficient amount of dialogues is available, conclusions can be drawn on whether accommodation generally occurs under specific conditions or not. Although this methodology produces informative results, there are two arguments against it: first, it has been argued whether this correlation is the result of accommodation or not. The alternative explanation provided, is that it may be a result of topic liveliness (Benus 2009), or of the overall liveliness of the dialogue (Bosch *et al.* 2005). Second, it fails to capture the dynamic evolution of accommodation over time as the dialogue progresses (Edlund *et al.* 2009).

Within-dialogue measurements can also be sub-categorized into continuous and non-continuous methods. Continuous methods consider utterances, turns, or other arbitrarily constructed units, on which a feature value can be measured or accumulated (averaged). These values are then located on a single point of the dialogue time-line. For example, the "center" of the utterance was used in (Nishimura *et al.* 2008), or a particular recurring syllable was used in (Kakita 1996). This process results in a time series for each speaker. Another option for creating a time series is to use the values from one speaker and linearly interpolated values from the second speaker at these points (Nishimura *et al.* 2008; Edlund *et al.* 2009). These time series are often simply inspected, in order to provide preliminary evidence of dynamic patterns (Kakita 1996; McRoberts and Best 1997; Campbell 2009). In other cases, the time series undergo statistical analysis, with one of various methods available in standard statistics handbooks (e.g. Chatfield 1996).

METHODOLOGY	FEATURE	CORPUS	STUDY
Time series (lag regression)	Rhythm, duration coordination	Mother-infant	Jaffe et al (2001)
Across dialogues & Time series (plot observations)	F0 accommodation	Parent-infant	McRoberts and Best (1997)
GLMM, frames of fixed length after prime	Syntactic priming	Spontaneous Task-oriented	Reitter et al (2006)
ANOVA, perceptual test of pronunciation pre-task, task, post-task	Phonetic convergence	Task-based	Pardo (2006)
Time series (trend line fit)	F0 convergence & divergence	Scripted answer-question pairs	Kakita (1996)
Across dialogues & Histograms	Pause duration overlaps	Spontaneous face-to-face & telephone	Bosch et al (2004, 2004b, 2005)
Time series (spectral analysis)	F0 and Intensity synchrony	Laboratory Adult conversations	Buder & Eriksson (1997, 1999)
Histograms & phase component	Syllable & accent timing entrainment	Spontaneous Elicitation	Benus (2009)
Superimposed time series plot observations	Multimodal synchrony	Multi-party conversation (video)	Campbell (2009)
Time series (recurrence analysis)	Swing & eye movement entrainment	Task oriented	Richardson et al (2008)
Time series (by interpolation) Pearson coefficient	Pause and gap length accommodation	spontaneous	Edlund et al (2009)
Linear regression, frames of fixed length after prime	F0 & lexical convergence	Tutorial sessions	Ward & Litman (2007,2007b)
Across dialogues	Speech rate adaptation	Task-oriented (telephone)	Ward & Nakagawa (2004)
Time series (by interpolation) lag zero coefficient	F0, Intensity and speed synchrony	Spontaneous	Nishimura et al (2008)
Percentage of success	Lexical entrainment	Spontaneous & WoZ & text	Brennan (1996)
Same word/different word ratio	Lexical entrainment	WoZ – Automatic translation	Fais (1996)
Across dialogues, Half-split dialogue ANOVA	F0, Intensity, speech rate, pause length	WoZ – Multimodal SDS	Oviatt et al (2002, 2002b,2004)
Per turn type ANOVA	Speech rate adaptation	WoZ – Multimodal SDS	Bell et al (2003)
Half-split dialogues t-test	Intensity, speech rate	WoZ – Quiz SDS	Suzuki & Katagiri (2003, 2004, 2005)

Table 4.1: Measurements of inter-speaker accommodation in various studies

The advantages of continuous (time series) methods are that (a) the variations in the feature value over time are captured, hence analysis can be performed on a *single* dialogue (McRoberts and Best 1997), and (b) that it is possible to determine whether only one or both speakers converge/diverge (Jaffe *et al.* 2001). In addition, it is possible to identify cyclical patterns to which it is possible to fit

models based on their periodicity (Buder and Eriksson 1997; 1999). In the latter study, a physical function was given to the period of the fitted sinusoids, namely that of rhythmic entrainment across the two speakers during turn exchanges (different periods were found for F0 and intensity). Similarly, (Jaffe *et al.* 2001) proposed an "optimal lag" which was found to be the most significant in a series of lagged regressions between the two time series. (Jaffe *et al.* 2001) proposed that this may be evidence of rhythm (periodicity) in dialogue interaction. Aside from the question whether such assumptions are valid or not, the findings themselves are proof that continuous approaches reveal much more information about accommodation than across-dialogue comparisons. The disadvantages of time series methods are the increased complexity (Edlund *et al.* 2009), and the fact that the usual assumptions for time-series analysis (stationarity, normal distribution of variance) are probably not satisfied in a strict sense (this is discussed in section 7.4.1).

Non-continuous methods encompass all other within-dialogue measurements: priming measurements, for example, make use of fixed-length frames that are defined by the location of the prime. Histograms display the distribution of values for a feature (such as pause duration), which can often provide valuable information. A somewhat crude method of measuring within-dialogue accommodation is the "half-split" approach: a dialogue is divided into two halves of equal length, and a feature average (for each speaker) is calculated for each half (e.g. Oviatt *et al.* 2004). This can be used to show whether speakers converged, diverged, or not. Although this method has been criticized for the same reasons as across-dialogue approaches (Edlund *et al.* 2009), it does combine merits from both, as the result is, in a sense, a two-point time series. One can imagine further splits into quarters etc, but there is a trade-off: unless the "pieces" are big enough, the average of a calculated feature may be biased by local events in the interaction.

In conclusion, time-series is the only analysis method which has been used so far to measure inter-speaker accommodation in a continuous way. Despite the disadvantages that were mentioned above, time series analysis provides the most complete description of accommodation phenomena and constitutes the most promising route towards a quantitative model that can be useful for SDS, as online monitoring and real-time accommodation pre-require a continuous description.

5 Review conclusions

5.1 Motivation for investigating accommodation phenomena

Inter-speaker accommodation is a ubiquitous phenomenon in human interaction, which has been studied in various disciplines and has been explained in various ways. It covers a wide spectrum of phenomena, which encompass the entirety of communication channels: lexical and syntactic choice, pronunciation, prosodic features, rhythm, posture, gaze and movement are the modalities along which interlocutors align their behavioural patterns. As highlighted in chapters 2 and 3, incorporation of methods to allow for "realistic" accommodation would significantly benefit spoken dialogue systems in a number of ways:

(a) Accommodation phenomena have been associated with smoothness of dialogue (Buder and Eriksson 1999) and communication efficiency (Pickering and Garrod 2004). Therefore, SDS that display such behaviour would be more efficient in communicating with the user. This should not be confused with efficiency of task completion in terms of dialogue duration, or any similar measure. Assuming that an SDS is designed for "human-like" dialogue, it *should* be able to communicate more efficiently if accommodation was built-in.

(b) According to Communication Accommodation Theory (Giles *et al.* 1987), convergence to the interlocutor's speech is evaluated positively if the situational constraints do not dictate otherwise. In other words, it is natural for interlocutors to converge, due to similarity attraction. Therefore, an SDS implementing the human metaphor could exploit convergence in order to make the interaction more pleasant for the user.

(c) As has already been mentioned, accommodation is a ubiquitous phenomenon in human speech, even if people do not consciously realize it. Consequently, an SDS that exhibits accommodating behaviour is likely to be perceived as more natural (or human-like), enhancing the "human metaphor", as proposed in (Edlund *et al.* 2008).

(d) Prosodic modeling for speech synthesis may directly benefit from a/p feature convergence models. Traditional prosodic models that have been developed for monologue speech have faced the mapping problem (see section 2.4.1), which is the transformation of a prosodic representation to an actual prosodic contour. Typically, these realizations of the abstract prosodic representations have a constant baseline, which is considered as speaker-specific (Dutoit 1997; Tatham and Morton 2005). If prosodic accommodation is taken into account, more appropriate realizations can perhaps be found, due to a baseline change which is consistent with the running dialogue. The resulting synthesized prosodic contours are likely to be perceived as more natural-sounding when considered in the dialogue context. The same

applies for prosodic modeling beyond fundamental frequency. For example, silence and filled pause duration modeling (Zellner 1994) for speech synthesis could benefit from adjusting predicted silence durations for inter-speaker accommodation.

(e) Classification of dialogue acts, both on-line and off-line is based on lexical, prosodic, syntactic and semantic/pragmatic information. Accommodation along any of these dimensions can inform this classification. For example, prosodic information is used to classify backchannelling expressions based on their pitch and duration (e.g. Rangarajan *et al.* 2007). The accuracy rates of the classifiers could be improved by changing parameter values according to the on-going pitch/duration accommodation in a given dialogue.

(f) Emotion recognition also relies heavily on prosodic correlates in the speech signal. Similarly to (e) above, the classification could be informed with accommodation information that is dynamically defined during the interaction.

(g) ASR typically shows high word error rates when then speech input is too variant. If speakers can be encouraged to adapt properties of their voice (such as speech rate, loudness) within certain limits, then ASR performance could be improved, as proposed in (Bell *et al.* 2003)

The above list of benefits is not inclusive, as it mostly focuses on accommodation of prosodic and temporal features of speech. The motivation behind investigating these particular features, apart from the potential benefits presented above, was already discussed in section 2.5: prosody has been perhaps the major "avenue" of improving on naturalness of synthesized speech. The problem of re-defining existing models that account for linguistic and para-linguistic variations of a/p features in a dialogue context has already been widely acknowledged (Mushin *et al.* 2003; Kohler 2004; Lee and Narayanan 2005). Temporal features, such as the duration of silences between dialogue utterances and the occurrence of overlapping speech are also inadequately dealt with in current SDS implementations (Raux and Eskenazi 2008). Investigation of accommodation phenomena related to both prosodic and temporal features constitutes a step away from speaker-listener studies and towards dialogue-based approaches to modeling these features.

Similarly, accommodation phenomena in other modalities are equally essential to developing human-machine interaction that can be perceived as human-like: lexical and syntactic choice, pronunciation, rhythm, posture, gaze and movement offer additional possibilities for improving on multimodal human-machine interaction. A replication of the entire range of this phenomena in the context of SDS would enhance the human metaphor significantly.

However, incorporating inter-speaker accommodation in human-machine interaction requires a quantitative description of these phenomena in order to replicate the behaviour adequately. As highlighted in section 4.6, past research has not accomplished this goal. The following section discusses some of the limitations of previously proposed measurements of inter-speaker accommodation, in relation to developing SDS that exhibit such behaviour.

5.2 Limitations to quantifying accommodation

A review of theoretical perspectives of inter-speaker accommodation phenomena was given in chapter 3. As highlighted in section 3.6, the majority of these theoretical models are based on positive empirical evidence acquired in laboratory conditions. The presence or absence of accommodating behaviour in some cases has been assessed by perceptual "expert" judgements, while little emphasis has been put on measuring the magnitude of these phenomena. In contrast, theoretical models have focused on the cause and function of accommodation. Such functions are cognitive alignment, communication efficiency, satisfaction of emotional needs, social approval or balancing a dyadic relationship. Several of these functions are relevant in the context of SDS, but without a quantification of the observations, it is impossible to develop systems that can replicate the behaviour observed in human dialogues. This problem was identified in (Oviatt *et al.* 2004):

> "One weakness of past research on interpersonal linguistic adaptation has been its lack of follow-through on quantitative research and user modeling. Instead, this literature has focused on qualitative descriptions of the social dynamics and context involved in linguistic accommodation. It has also relied on global correlational measures to demonstrate linguistic accommodation between two interlocutors. In future research, more quantitative predictive modeling will be needed on the process of linguistic convergence, including the magnitude and rate of adaptation of different linguistic features, the factors that drive dynamic adaptation and re-adaptation during human-computer conversation, and other key issues. Such models will be valuable in guiding the design of future conversational interfaces and their adaptive processing capabilities."[12], (Oviatt *et al.* 2004)

As noted in chapter 4, the mechanisms currently available for monitoring and quantifying accommodation are unsuitable for SDS that aim to mimic human-like interaction. Existing approaches to measuring accommodation are almost exclusively – with few exceptions – statistical. The typical process comprises (a) acquiring speech recordings, (b) extracting features and (c) performing statistical analysis or – in some cases – signal processing techniques in order to validate

12 (Oviatt *et al.* 2004) uses the term "linguistic" to signify any property of spoken language. The features studied in the same text are amplitude, speech rate and response latency.

the hypothesis of accommodation, or to compare the results of two or more experimental conditions. In assessing the limitations of existing studies of measuring accommodation, these three stages are discussed in the remainder of this section.

As highlighted in section 2.5, proponents of human-like SDS (Carlson *et al.* 2006; Edlund *et al.* 2008) have emphasized the need for investigating human behaviour in dialogues of spontaneous speech. The reason for this requirement is that spontaneous speech is human speech in its most natural form. Therefore, knowledge derived from investigating such corpora is more likely to perceived as "natural" when applied to SDS. Wizard-of-Oz SDS environments simulating application tasks can also be used, but care has to be taken that properties of natural human speech are not masked by the experimental constraints. Accommodation, in particular, has been found to be affected by task complexity (Pardo 2006) and talker role (Fais 1996) among other factors.

However, few of the studies reviewed in chapter 4 have used spontaneous speech in their investigation of accommodation phenomena (see Table 4.1). Some of the studies have used scripted dialogues, which were designed so that features could be extracted from identical lexical elements (Kakita 1996), or utterance types (Suzuki and Katagiri 2005). Despite the advantages of this approach in relation to robust feature extraction, the "dialogue" is artificial and the results of these studies cannot be generalized. A second group of studies used simulated human-machine interaction scenarios, in which subjects had the role of the "user" (Bell *et al.* 2003; Suzuki and Katagiri 2004; Oviatt *et al.* 2004). While these studies provided evidence of user accommodation towards the "system", it is doubtful whether they can be helpful in comparing human-human and human-machine interaction in this regard and informing improvements on the human-likeness of SDS. A third group of studies reported using spontaneous speech recordings (Brennan 1996; Bosch *et al.* 2004b; Reitter *et al.* 2006; Nishimura *et al.* 2008; Campbell 2009; Edlund *et al.* 2009; Benus 2009). However, as discussed in section 2.5, acquiring recordings of genuine spontaneous speech is not trivial, and careful consideration is required in order to record such dialogues.

The stage of feature extraction is also typically accompanied by a number of assumptions. Turns, in particular, are typically defined using an arbitrary turn attribution scheme (see section 2.3.2) which assumes speakers are holding and releasing the floor at specific points. However, such schemes are not adequate in describing spontaneous speech and thus introduce bias in the subsequent analysis. Another assumption commonly found is to extract features from entire utterances and "tie" them to a specific time point, such as the beginning (Kakita 1996) or the middle (Nishimura *et al.* 2008) of the utterance. While such conventions are convenient, they are not necessarily consistent with the process of speech production and perception in human speech: the prosodic realization of an utterance is not pre-determined before vocalization, but comes as a result of articulation effort (Xu

2005) and simultaneous feedback from the interlocutor (Heylen 2009).

Finally, statistical validation of inter-speaker accommodation has been accomplished in a variety of ways, but most of these methods are not helpful for quantifying/modeling this behaviour for SDS. A characteristic example is across-dialogue comparisons (Coulston *et al.* 2002; Bosch *et al.* 2004a; Suzuki and Katagiri 2004; Ward and Nakagawa 2004), in which subjects' speech features are compared across two or more different conditions; in some cases, the dialogue is arbitrarily split into two halves (Darves and Oviatt 2002; Suzuki and Katagiri 2005), resulting in a comparison between the first and second half; and yet it is clear, from any of the theoretical descriptions, that accommodation phenomena are *dynamic*: they (are thought to) evolve through the interaction and characterize it in terms of "coordination" or "synchrony". This can only be indicated by using a continuous measurement methodology, sampling at regular intervals or identified instances (depending on the features studied), in order to arrive at a model which describes the variations of these features that occur as a result of inter-speaker accommodation. Such a model can then be used in SDS in order to continuously monitor the user's speech (or other modalities) and adapt the system voice accordingly.

A promising approach in this direction is time-series analysis, which has been used in a number of studies reviewed in chapter 4. However, time series analysis is characterized by complexity, which discourages wide adoption of this technique (Edlund *et al.* 2009). Thus, several studies are limited to inferring conclusions by simply inspecting the time series plots (McRoberts and Best 1997), while a few take the next step and employ an analytical approach (Buder and Eriksson 1999; Jaffe *et al.* 2001; Nishimura *et al.* 2008; Richardson *et al.* 2008). However, only one of these proposed a model for monitoring user accommodation and adapting the system voice to accommodate to that of the user (Nishimura *et al.* 2008).

The problem of quantification is perhaps most evident in studying accommodation of temporal features, such as the duration of silences before/after utterances. The phenomenon is studied from two distinct viewpoints: Communication Accommodation Theory (Giles *et al.* 1992) proposes that this is another form of socially-driven behaviour, while studies on *rhythmic entrainment* (Jaffe and Feldstein 1970; Wilson and Wilson 2005) suggest that interlocutors are rhythmically "coupled" when engaged in dialogue. Evidence is weak for both: across-dialogue comparison of silence duration convergence among speakers (Bosch *et al.* 2004b) does not constitute solid evidence, as it can be attributed to other causes, such as dialogue or topic liveliness (Bosch *et al.* 2005; Benus 2009); turn-based time series approaches show partial evidence: only a portion of the dialogues exhibit simultaneous variation of silence duration among speakers (Edlund *et al.* 2009); and there is little empirical support for "coupling" theories (Benus 2009). Therefore, temporal accommodation

is a subjectively observed phenomenon, but there is weak evidence for it, especially in the case of spontaneous speech.

It is evident from the review that inter-speaker accommodation phenomena have not been described adequately in respect to their manifestation; and this is a significant obstacle towards their implementation in SDS. Therefore, further investigation of the form of accommodation is required, in order to extract information that can be useful for SDS.

5.3 Conclusion

Inter-speaker accommodation offers the potential of improving on naturalness and efficiency of SDS in various ways. The theoretical foundations and experimental findings support this view. In particular, prosody and the temporal structure of dialogue are the most promising features of human dialogue which would arguably improve on the naturalness of SDS the most. In addition, speech synthesis technology allows for straightforward manipulation of these features, thus making incorporation of inter-speaker accommodation in SDS feasible, provided that an adequate model exists.

However, existing methods of measuring inter-speaker accommodation have not adequately quantified these phenomena, and have also been based on assumptions which are inconsistent with naturally occurring human speech. Therefore, an investigation of these features in spontaneous human dialogues is required, as this type of interaction is the most general case of spoken communication and allows inference of knowledge without making assumptions on the possible effects of arbitrarily imposed constraints. A methodology for acquiring high audio quality recordings of spontaneous speech is presented in chapter 6.

In chapter 7, a methodology for quantifying/monitoring accommodation is presented which deals with these limitations by considering a frame-based representation of the dialogue: features are extracted from each speaker's utterances as averages of overlapping frames of fixed length, thus circumventing the requirement to define turn-exchanging points. This process, termed TAMA (Time-Aligned Moving Average), results in two time series (one per speaker) in which the time indices for both speakers are the same. This enables the consideration of a dialogue as a bi-variate process which demonstrates *feedback,* as shown by the statistical analysis. The magnitude of accommodation can be estimated by statistical modeling, which allows for direct implementation of accommodating behaviour in an SDS environment. A first approach towards the latter goal is demonstrated in chapter 9.

Chapter 8 presents an investigation of accommodation of temporal features. Due to issues of data

sparsity and variation introduced by the discourse structure, the TAMA methodology is not adequate in itself to describe accommodation phenomena of temporal features. Thus, an additional novel dialogue representation is presented in the same chapter, which explores the effect of turn "shares" on the variations of temporal features. Turn shares represent the "floor balance" in a dialogue over time, i.e. whether the floor is shared or dominated by either speaker. The proposed representation provides additional evidence of temporal accommodation to that provided by across dialogue comparisons (Bosch *et al.* 2004b), turn-based time series approaches (Edlund *et al.* 2009) and TAMA.

6 Design of research methodology and data acquisition

6.1 Overview

Following the discussion in the previous chapter and the findings in the literature review, the overall aim established is to formulate a continuous quantitative description of inter-speaker accommodation (of a/p and temporal features), based on analysis of recorded spontaneous dialogues. This chapter describes the overall research design (section 6.2), as well as the design and implementation of the audio recording environment (section 6.3) and experimental scenarios for eliciting spontaneous dialogues (section 6.4). Section 6.5 describes the procedures followed for annotation of the corpus and extraction of prosodic and temporal features which are analyzed in later chapters.

6.2 Research design

The research methodology was designed according to a specification that is described here (see Table 6.1). The overall goals were:

(a) acquisition of high audio quality recordings of spontaneous dialogue speech, for the purposes of this work, but also beneficial for future research,

(b) analysis of the recordings for evidence of inter-speaker accommodation in acoustic/prosodic and temporal features,

(c) formulation of a quantitative description of inter-speaker accommodation, and

(d) proposal of methods which can utilize inter-speaker accommodation in spoken dialogue systems.

Each of the above main goals is divided into secondary objectives. For example, (a) above required both a recording environment and an experimental design, in order to elicit spontaneous speech from the participants. Taking into account the audio quality issues discussed in section 2.5, it was decided that (1) CD quality (44.1 KHz, 16-bit) would be the absolute minimum quality for the recordings, (2) since dialogue recordings are needed, a two-channel approach would be the most efficient, and (3) a separate soundproof environment for each speaker (to avoid cross-channel noise contamination) would be best. As was also discussed in section 2.5 mood induction procedures (MIPS) were considered as the best method for spontaneous speech elicitation in laboratory conditions. In addition, unconstrained dialogues (without the MIP method) between subjects were also considered, as this method of obtaining spontaneous speech has also been proposed by several studies, as discussed in section 4.6).

Similarly, (b) can be considered as a two-stage process: the first step is feature extraction, which essentially is taking measurements of relevant properties from the speech signal. As discussed in chapter 5, acoustic/prosodic (a/p) and temporal features were identified as the most relevant for improving naturalness in SDS. In particular, these are pitch (F0), pitch range, speech rate, intensity, inter-speaker silence duration and occurrence of overlapping speech. The second step is the subsequent analysis of the extracted features, in this case for the purpose of describing the phenomenon of inter-speaker accommodation. As was pointed in section 4.6, and also by others (Edlund *et al.* 2009), only a small number of studies have considered a *continuous* approach to describing the phenomena, although this approach is the most promising in terms of usability of the results. Thus, this methodology was seen as the most suitable for investigation of the phenomena.

Main Objective	Requirement	Specification
Recordings of spontaneous dialogues	Channels	2 (in separate soundproof environments)
	Audio quality	CD (44.1 KHz/16-bit) or better
	Spontaneous speech elicitation	MIPS task-based experiments unconstrained speech
Analysis of corpus	Feature Extraction	Prosodic and Temporal Features pitch, intensity, speech rate, pause duration
	Main analysis	Continuous – time series approach
Description of Accommodation	Quantitative	Statistic evaluation per dialogue and per individual feature
	Bi-directionality & feedback	
SDS implementations	Simulations & model fitting	Off-line manipulation
	Test platform	Wizard of Oz

Table 6.1: Specification of the overall research methodology

A quantitative description of inter-speaker accommodation must take into account the theoretical predictions described in chapter 3. More specifically, the influence of each speaker's prosodic and temporal properties of speech on the respective properties of the other, is considered as a first step towards this description. This is schematized as shown in Figure 6.1 below.

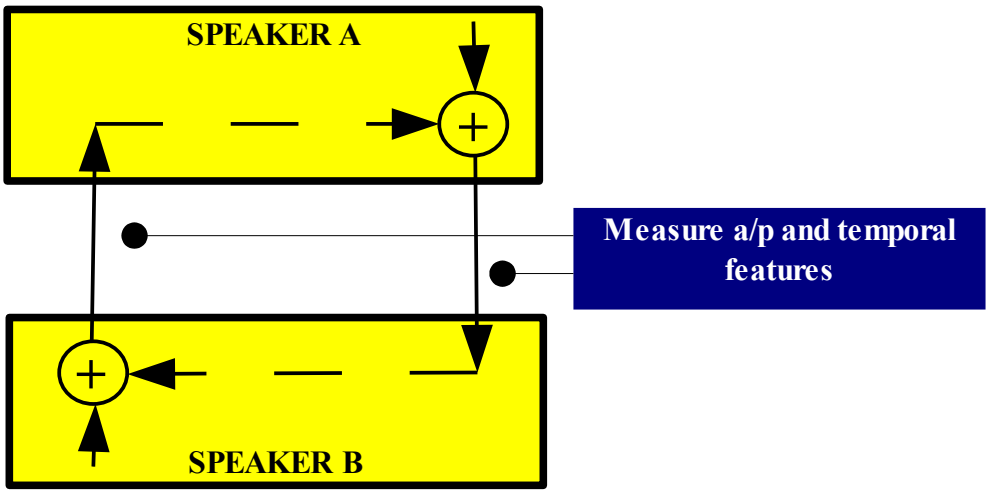


Figure 6.1: Schema for describing measurement of inter-speaker accommodation of speech features

The schema shown is a representation of a dialogue between speakers A and B. This representation assumes that any utterance from speaker A (downward vertical arrow on the right) is *perceived* by speaker B (left-pointing dashed arrow at the bottom) in a way that influences B's own utterance (upward vertical arrow on the left) and vice versa. The summation symbols (+) denote that there is an influence both by the each speaker's own (inherent) speech properties, as well as those of the interlocutor (therefore a "summation"). The yellow rectangles denote internal processes (speech perception and production) of either speaker, while the white space in-between denotes the external (shared) environment.

The result of the summation is the actual, uttered speech which can be recorded and analyzed. This schema hypothesizes that there is a *feedback* loop involved in the process of dialogue. The goal is therefore to evaluate this hypothesis, by quantifying the influence of each speaker's speech properties (a/p and temporal) on the actual (measured) properties. Further, if *both* speakers influence each other, then accommodation is bidirectional. In case one of the speakers is not influenced by the other, then the above schema is simplified to an open-loop system and accommodation is uni-directional.

Finally, as suggested in (Edlund *et al.* 2008), the most prominent methods of evaluating human-machine interaction against human dialogues are those of data manipulation (off and on-line) and Wizard of Oz experiments (see section 2.2.4). Thus, both of these evaluation methods were planned at the beginning of this project. The overall design of the methodology is shown in Figure 6.2 below.

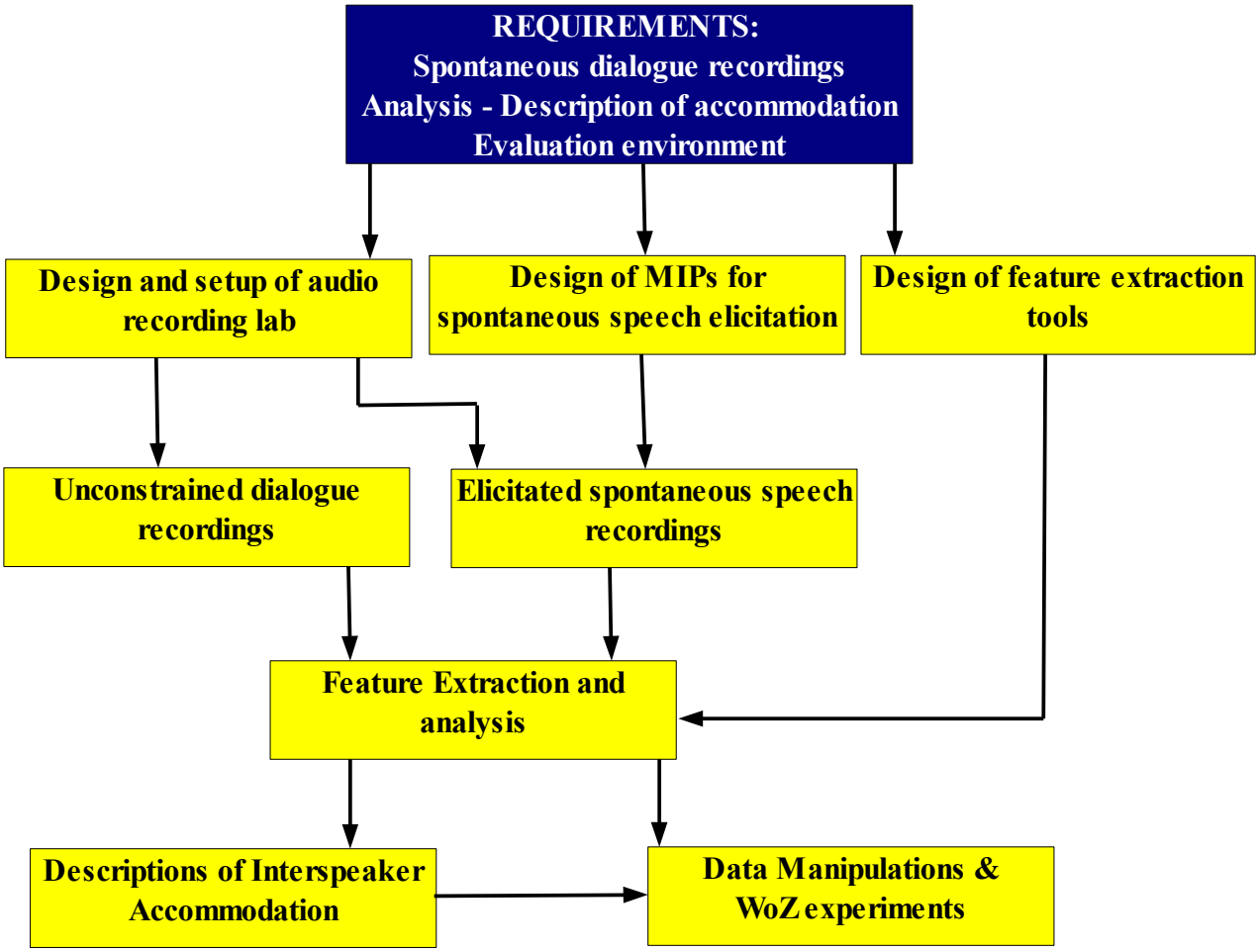


Figure 6.2: Outline of methodology in block diagram form

6.3 Audio recording environment

In order to ensure noiseless and optimal quality recording of dialogues, the audio recording environment comprised two soundproof isolation booths (see Figure 6.3). The standard equipment for each booth comprised a 17" flat monitor, a Beyer DT150® headphone set and a Neumann U87® microphone. The audio equipment was connected to a ProTools HD® console, and controlled by an Apple Mac Pro® workstation running DigiDesign ProTools® audio processing software. These internal monitors were connected to separate external workstations or game consoles depending on the experiment

The advantages of this setup are (a) soundproofing of the booths only is required (rather than an entire room), (b) each speaker is recorded in a *separate* audio channel, (c) subjects are not situated in a such a way that they might feel "being watched", and (d) visual feedback can be introduced by use of cameras.

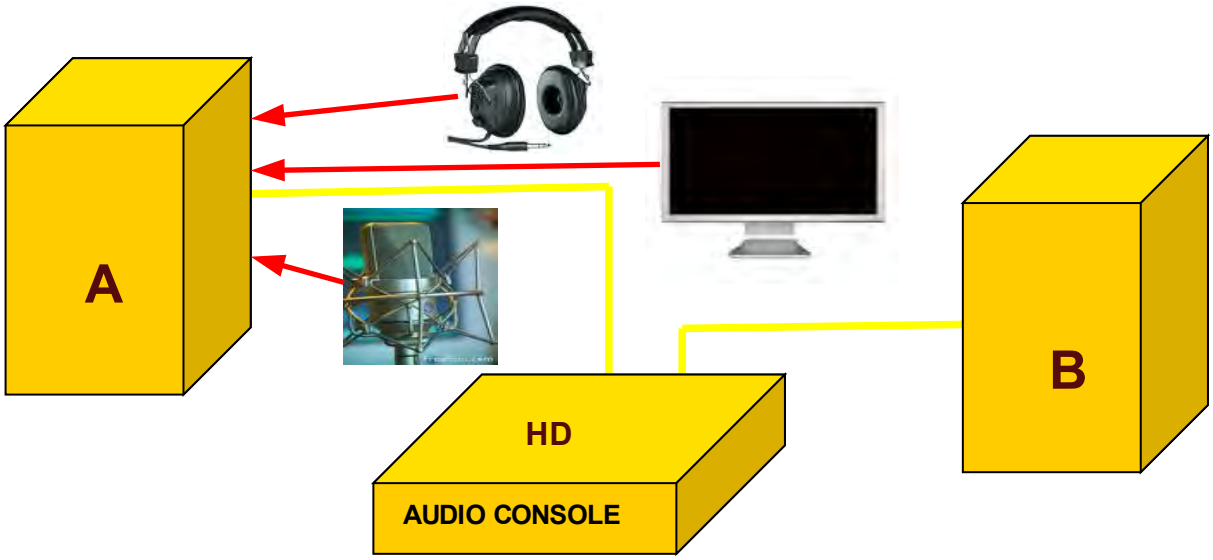


Figure 6.3: Schematic of audio recording setup

Clearly, (a) above is mainly a resource consideration, although there are significant differences between the two options: the booths were placed on wooden frames, thus were not in direct contact to the laboratory floor (during installation, this was found to reduce noise caused by floor vibrations due to footsteps etc). They were soundproofed with foam material on all 6 faces, including I/O cable outlets. LED light-chains were installed inside each booth for lighting. The flat panel monitors, speakers and microphones were connected to their inputs via long cables, thus moving the workstations at a sufficient distance from the booths and ensuring no interference from noisy computer components, such as cooling fans All of the above installations ensured a low-noise environment inside the booths.

Recording both (or more) subjects in the same space would require the use of close-talk or contact microphones. In the first case, obtaining a high signal-to-noise ratio without cross-channel contamination is difficult. In other words, both speakers would be recorded on both channels, although the amplitude of speaker "A" on channel "B" would be much smaller than that of speaker "B" and vice versa. Signal sources can be separated in each channel, using audio *source separation* (Persia *et al.* 2007), but the signal distortion (artefacts) introduced by this process were deemed inappropriate for the purposes of analysis and re-usability of the corpus, or at least unnecessary if they could be avoided.

Contact microphones all but extinguish this problem, but are known to produce a "tinny" sound, due to the sound signal being transmitted through bone tissue, which results in attenuation of some

frequencies. Thus, they are more suitable for voice activity detection (e.g. Jaffe *et al.* 2001) or fundamental frequency measurement and glottal source estimation by *inverse filtering* (Askenfelt *et al.* 1980; Walker and Murphy 2007).

Consequently, the chosen method (separate isolation booths) ensures the best possible recording of each speaker in a separate audio channel.

The booth itself was considered to provide some "privacy" to the subjects, as they cannot be seen from outside. This was thought to encourage spontaneous behaviour in the case of task-based mood-induction experiments, as the presence of other people in the setting (e.g. experimenters) might bias the subjects' responses, due to the feeling of being watched. This view is supported by related research studies (Gross and Levenson 1995; Fernandez and Picard 2000; Picard *et al.* 2001) which propose a relaxed and isolated environment for inducing spontaneous speech.

The drawback, of course, is that direct visual feedback is not possible, and thus face-to-face conversations could not be recorded. The possibility of using cameras to enable facial communication was deemed sufficient to overcome this problem, considering the fact that the goal of studying a/p and temporal features did not require visual contact: several other relevant studies (see chapter 4) have used corpora comprising telephone conversations. However, and particularly in relation to inter-speaker accommodation phenomena, it has also been found that *relayed* visual feedback is *not* equivalent to face-to-face communication (Richardson *et al.* 2008).

In conclusion, the particular setup was chosen for providing the best possible audio quality and a suitable environment for recording spontaneous dialogues. The recording console and equipment made possible the recording of audio at a sampling rate of 192 KHz /24-bit (in lossless WAV format), which was used for all experiments, while the soundproof booths provided for a low noise environment and a separate audio channel for each speaker[13].

6.4 Recording experiments

Two types of recording experiments have been used in the work described in this dissertation. The first type is unsolicited, unconstrained dialogues that were recorded with subjects situated in the booths. The second type is spontaneous dialogue recordings elicited by mood induction procedures. Both types are discussed in the next two sections (6.4.1 and 6.4.2). Detailed information of the dialogues can be found in appendix A.

13 The installation of the described audio recording laboratory was a collaborative undertaking within the SALERO project (www.salero.info), which was funded by the EU. The laboratory has been used for other projects, such as the acquisition and annotation of an emotional speech corpus (Cullen 2008a).

6.4.1 Unconstrained dialogues

These dialogues were recorded while loosely acquainted or well-acquainted subjects (mostly DMC[14] and DIT staff and students) conversed in pairs from within the isolation booths. These conversations were primarily recorded for the purposes of a language learning research project called FLUENT[15]. In FLUENT, these recordings aim to provide a non-native language learner with audio material from native speakers, in three gradually "ascending" stages: (1) short, scripted conversations, (2) "role-playing" dialogues (such as ticket-booking), and (3) unconstrained dialogues. Dialogues acquired with method (3) were selected for analysis of inter-speaker accommodation, based on a quality rating given by the FLUENT research group to each dialogue. The dialogues comprise unconstrained speech that is not organized in any way. The dialogues can be characterized as "friendly chats". There are many unpredictable topic changes, and there is a fair amount of spontaneous dialogue acts (interruptions, laughter, disfluencies, repairs), which would classify this speech as spontaneous. Therefore, these dialogues were considered as appropriate for studying inter-speaker accommodation.

However, it could be argued that subjects in these experiments were not as "relaxed" as they would be in a real-life setting, due to the presence of the recording equipment and the awareness of being recorded. In order to overcome this, one needs to turn towards experimental settings that require subjects to participate in a task, as task requirements are found to distract speakers from the recording setting and communicate more freely (Gross and Levenson 1995; Fernandez and Picard 2000; Picard *et al.* 2001).

6.4.2 Elicited spontaneity

A variety of experimental scenarios for eliciting spontaneous speech were considered in the design phase (Vaughan *et al.* 2006; 2007). These were primarily designed to elicit human emotions. However, since the chosen method of emotion elicitation was to encourage spontaneous speech, these scenarios were considered for analysis of inter-speaker accommodation.

The first of these experimental designs was a LEGO® puzzle which has also been used in (Kehrein 2002). In this scenario, one of the subjects is given the instructions for constructing an object (in this case a fire engine), while the other subject is given the LEGO pieces. In the simplest case, this encourages the two subjects (who are situated in the two separate booths and have no visual contact to each other) to get involved in the construction of the puzzle, a process which provides for natural

14 Digital Media Center, www.dmc.dit.ie
15 FLUENT is a language learning project, funded by Enterprise Ireland (http://www.dmc.dit.ie/2006/projects.html)

interaction between the participants. An extension to this idea is to provide the subjects with fewer pieces and/or misleading instructions, which is more tailored to the idea of inducing mild frustration, for the purpose of recording spontaneous emotions (Cullen *et al.* 2006). An important point in this scenario - from an accommodation point of view - is that the two subjects have distinct roles (information giver vs information receiver). While this is perhaps also relevant from an SDS application point of view, it was considered that – for the purpose of studying inter-speaker accommodation – any task should be "symmetrical" for the two subjects.

Another proposed scenario was that of a dice game known as "Mexican" or "Bluff" in which players roll two dice in turns. Each player has to claim a roll higher than the opponent's previous roll. If a bluff is called then that player loses a "life", while if the roll was actually the one claimed, then the player who called the bluff loses a life. While this scenario is symmetrical and also suitable for acquiring spontaneous emotions, it was considered that the lexical variety in the corpus would be small (mainly digits that describe rolls) and that the game itself has a short duration with only two players, unless they are given a large amount of lives, in which case it becomes very repetitive.

A third idea, proposed in (Johnstone 1996), was to record subjects while they were playing a computer game (Gears of War®[16] - a combat-style game). Actual sessions were recorded using this method. This required the additional installation of two Microsoft XBOX II ® gaming consoles, which were connected to the monitors in the booths. The subjects were playing in the same game area (via LAN connection) and had to combat each other in-game. Although this method is suitable for obtaining spontaneous emotions, it is less suitable for obtaining spontaneous *conversation,* since the subjects tended to remain silent for long periods of time. Most of the speech material occurred in "bursts" along with laughter or other non-verbal expressions, typically when a significant event happened in game. Minimal conversations occurred that were sparse and of very short duration. Thus, these recordings were not used in the study of inter-speaker accommodation.

6.4.3 The "shipwrecked" scenario

The experience from the early efforts described in the previous section led to the conclusion that the experimental design should comprise a task for the subjects to be involved with, while having a number of desired properties (a) it must require *discussion,* thus encouraging spontaneous conversation, (b) it must be symmetrical, i.e. experienced equally by both participants, (c) it must not constrain the subjects to any specific linguistic content (as in the case of the dice game), and (d) some motivation should be provided to the subjects to get involved with the task promptly.

16 http://gearsofwar.xbox.com/AgeGate.htm

Figure 6.4: Shipwrecked scenario

The above specification led to the design of the "shipwrecked" scenario (see Figure 6.4). In this experiment, the two subjects experienced a hypothetical shipwreck, from which they had to survive. In order to accomplish this, the two subjects had to agree on which items from those shown on-screen were the most essential and in what order. Thus they had to rank the 15 objects shown in the picture by order of importance in surviving the hazard. In addition, a time limit of ten minutes was imposed, so as to encourage quick involvement from the subjects. The result was that the conversations were relatively focused, thus eliminating the problem of long stretches of silence that was encountered in the computer game experiment. In an earlier version of the experiment the subjects were given a list of the objects on paper and a pen to write down the ranks. However, this was found to introduce noise in the recordings. The inclusion of pictures instead of object names required the subjects to name the objects themselves. Thus, the corpus can be used for investigating lexical accommodation, in addition to a/p and temporal features. Based on the same procedure, two more "hazard" scenarios were implemented: an expedition in the Himalayas, in which the subjects had lost their guide and path, and a space mission, where the subjects had to abandon their spaceship and get into a rescue pod. The task in both these cases was identical (ranking a set of 15 objects relevant to the task). These two sets of objects are shown in appendix A.

A further expansion of this experimental design was the inclusion of an on-line performance score. This score was automatically assigned and shown on-screen by an "intelligent" system, based on the "correct" ranking. This was actually a Wizard-of-Oz implementation, in which the changes in

the score shown were always the same regardless of the choices that the subjects made (there was no “correct” solution). The purpose of this was to record the subjects' reactions when they thought they were doing well with the task or when their score was dropping. Since this expansion did not alter the task and recording conditions significantly (in reality it only made the task appear more difficult), these recordings were also used in the study of inter-speaker accommodation.

Conclusively, a total of 30 dialogues were recorded using all methods, as shown in Table 6.2. The recording experiments for some categories are on-going: the table contains those dialogues that were analyzed for inter-speaker accommodation.

Method	Number of dialogues	Average Duration (min)	Total Duration (min)
Unconstrained	**8**	**20**	**161**
Shipwrecked	**14**	**8**	**108**
Shipwrecked + ranking score	**8**	**9**	**76**
Total	**30**	**-**	**345**

Table 6.2: Recorded dialogues

6.5 Corpus annotation and feature extraction

This section describes the annotation and feature extraction procedure followed in the analysis. There are three distinct steps in this procedure: (a) segmentation of the continuous recording into speech/silence, (b) annotation of non-silent segments with suitable labels and (c) feature extraction from the annotated segments. These three separate procedures are described in sections 6.5.1, 6.5.2, and 6.5.3 respectively.

6.5.1 Silence/ non-silence segmentation

The process of segmentation of a continuous audio stream into speech/silence segments is termed *chronography* (Lennes and Anttila 2002). The result of this process is typically a representation of the form shown in Figure 6.5, in which black and white areas denote speech activity and silence respectively (Lennes and Anttila 2002; Campbell 2009).

Segmentation can be performed either manually or automatically. In manual segmentation, a human annotator listens to the audio stream and demarcates the speech/silence areas one by one. This method produces adequately precise segmentation (±10ms) and, in addition, can be combined with

annotation of non-silent, non-speech areas. The latter step is explained in section 6.5.2. The disadvantage of manual segmentation is that is a repetitive and tedious process, which makes it costly and inefficient, especially for large corpora.

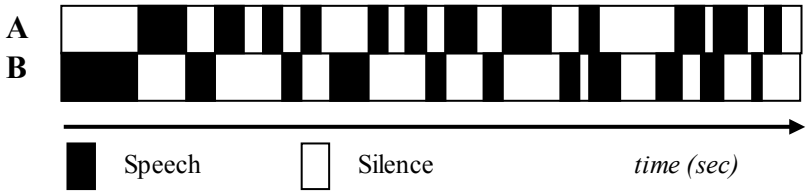


Figure 6.5: Chronographic represenation of dialogue (two speakers A, B)

Alternatively, an automatic segmentation can be achieved by means of a voice activation detection (VAD) algorithm. A simple implementation of such an algorithm is that of segmentation based on intensity and duration thresholds. As a first step, the audio stream is divided into frames, the length of which defines the "resolution" of the algorithm (e.g. ~10ms). The intensity is calculated for each frame based on Equation 2.1. Depending on the intensity relative to the intensity threshold, a frame is characterized as silent/non-silent. Adjacent silent/non-silent frames are joined together in silent/non-silent segments, respectively. This yields numerous segments that are shorter in duration than the minimum duration thresholds (which can be different for silent/non-silent intervals). As a last step, these segments are "erased" and neighbouring segments are joined. This has to be performed both for silent and non-silent intervals (in either order).

The above algorithm was implemented in the speech analysis software Praat (Boersma and Weenink 2009), originally using a Praat script[17] which is available on-line[18], and subsequently using a built-in command that was included in later versions of Praat (see appendix C).

The resulting segmentation using the automatic method typically contains errors. Areas that are non-silent may be annotated as speech and vice versa. This occurs because a "flat" intensity threshold cannot capture the possible variations in voice intensity throughout an entire dialogue. A high threshold "misses" utterances spoken much less loudly then average, while a low threshold captures too much extraneous noise, such as air stream from the mouth and nostrils when a subject is not speaking. A reasonable trade-off value can be found by manually adjusting the threshold value, but this cannot overcome all the problems. For example, stop-consonant (/p/ /k/ /t/) closures are typically cut-off from the speech segment and annotated as silence. Thus, manual corrections are again required for an adequately precise segmentation to be obtained. The resulting method, which was used for segmentation of all the dialogues in the corpora used in this thesis, is a *semi-*

17 Praat software operates as a shell where objects such as sounds can be queried or modified by means of commands. A series of commands can be executed as a shell script, also known as Praat script.

18 http://www.helsinki.fi/~lennes/praat-scripts/public/mark_pauses.praat (01/04/2010)

autonomous process: Automatic segmentation using the built-in Praat command, followed by manual correction of the output segments. An example segmentation using Praat and Mietta Lennes's script is shown in Figure 6.6 (silences marked by "xxx").

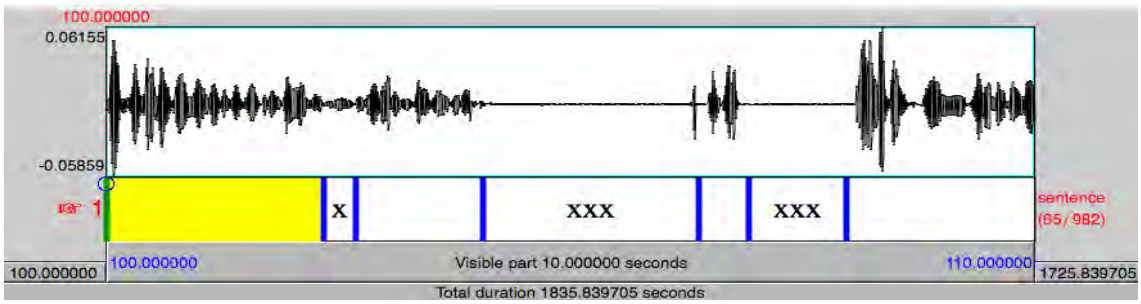


Figure 6.6: Segmentation of audio stream into silent/non-silent intervals

6.5.2 Annotation

This section describes the corpus annotation procedure followed in the work described in this dissertation. The output of the automatic segmentation process is a "textgrid" Praat object. This type of object is a time-line with marked boundaries, which define "intervals" (or segments). The timeline is shared between the sound object and textgrid object, in a way that boundaries mark silent and non-silent intervals, as shown in Figure 6.6. During the manual correction step that was described in section 6.5.1, the intervals are labeled for content according to the simple annotation schema shown in Table 6.3 below.

Label	Description
s	**Speech interval**
p	**Silent interval**
l	**laughter**
b	**Breathing noise**
n	**Other non-speech noise**

Table 6.3: Labels for annotation of textgrid intervals

The speech intervals, marked "s", denote any type of vocal activity by the speaker. This means that nonsense words, such as "uhm", "err", and filled pauses are considered as speech. This is justified from the point of view of further analysis. These utterances were observed to be prosodicaly similar to actual words (in the linguistic sense) and are thus further analyzed for prosodic features. Nonsense words, for example, frequently appear as back-channeling expressions in the corpus (both task-based and unconstrained). By comparison to "proper" lexical elements used as backchannels,

such as “yes”, it was found that these nonsense words serve the same purpose (acknowledgment of understanding/continuing attention) and exhibit similar prosodic structure. Nonsense words are not dictionary words, but the former are in all other ways *equivalent* to the latter: function, vocalization, prosodic structure. Filled pauses, which are typically of the form of elongated vowels, are also classified as speech, on the same premise as before: they represent vocal activity by the speaker and are prosodicaly similar to well-formed utterances, in terms of average pitch, intensity and pitch range. Therefore, it was decided that these should be treated as speech for the purpose of prosodic analysis. By extension, any prosodicaly “speech-like” interval uttered by the speakers was classified as speech, regardless of timing or function in the dialogue.

In contrast to the above rule, occurrences of laughter, marked “l”, were not classified as speech and were not prosodicaly analyzed. Laughter was common in all recorded dialogues. From a prosodic point of view, laughter is characterized by short repetitive bursts of high pitch and intensity, a pattern largely different from that of speech, which exhibits smoother pitch and intensity contours. In addition, pitch and intensity peaks fall outside their normal range during laughter. As these values introduce bias to the acoustic/prosodic analysis, it was decided to exclude them. Importantly, this did not apply to instances of “laughing” speech, which is audible speech uttered by a speaker who is laughing at the same time, but only to instances of pure laughter. The purpose of the distinct label is that laughs are still considered as “contributions” of the speaker, for the purposes of temporal analysis.

Similarly, the “b” and “n” labels denote breathing and other non-speech noises respectively. Breaths are quite common at the beginning of utterances and are often loud enough to be captured by the intensity-threshold algorithm. Due to their high intensity and non-voiced nature, breaths introduce bias to prosodic analysis and thus had to be located and labeled appropriately. As in the case of laughter, breaths were considered important for the purpose of temporal analysis. A long inhaling sound before an utterance may be signaling the intention to speak, and is therefore considered as a contribution by the speaker. The 'n' label groups together all other unvoiced, non-speech sounds (coughing, nasal inhalation, lip-smacking etc).

Silent intervals were annotated as pauses, marked “p”, and contain silence but also certain types of extraneous noise. This noise includes accidental knocks on the microphone stand or other surfaces that are “picked-up” by the intensity threshold algorithm. Such noises are not considered part of the interaction, and are thus not labeled. Instead, any interval that is automatically marked as non-silent because of extraneous noise was manually annotated as silent instead. This is significant mainly for the purposes of temporal analysis, as these noises are relatively infrequent and thus do not introduce bias in the prosodic analysis.

6.5.3 Feature extraction

Following segmentation and annotation of the audio files, feature extraction was carried out using the Praat software. The various steps described in this section were implemented as a collection of Praat scripts which can be found in appendix C.

As described in the previous section, the audio files were semi-automatically segmented and annotated with the labels shown in Table 6.3. Prosodic features were extracted using built-in Praat algorithms (Boersma and Weenink 2009) from intervals marked with the "s" label, henceforth termed speech intervals. The features measured on each speech interval were as follows:

(a) Fundamental frequency (F0), or pitch[19], was measured (in Hz) using the built-in Praat function that is based on the autocorrelation method (Boersma 1993). For each speech segment, the built-in function computes a pitch contour. Querying the pitch contour in the Praat environment yields a minimum, a maximum and an average value (arithmetic mean). The minimum and maximum were used to calculate pitch range. However, this method of pitch range calculation was too error-prone due to erroneous pitch values introduced by the algorithm, such as octave jumps or mistakenly calculating pitch values for non-voiced regions. Thus pitch range was consequently calculated as 2*std, the standard deviation of pitch, which can also be found by querying the pitch contour.

(b) Intensity, was measured (in dB) using the built-in Praat function that is based on Equation 2.1. For each speech segment, the built-in function computes an intensity contour. Querying the intensity contour yields a minimum, a maximum and an average value (arithmetic mean). However, the minimum and maximum were not used in further analysis. The built-in Praat function was used with the option "subtract mean" enabled. The purpose of this option is to subtract the "DC offset" introduced by audio recording equipment. Since the audio equipment used was of very high quality, with a signal-to-noise ratio greater than 90 dB, disabling the option yields negligible difference in the computed intensity values.

(c) Speech rate was measured (in vowels/minute) by counting the number of detected vowels and dividing by the length of the speech segment. This method yields only an approximation of speech rate (Pellegrino *et al.* 2004). However, since the purpose was to *compare* the speech rate of two speakers, the approximation was deemed sufficient in order to assess inter-speaker accommodation of speech rate. The vowel detection method used is based on calculating the derivative of the intensity contour (Press *et al.* 1992) and detecting vowel onsets and offsets based on steep rises, falls and peaks (Cummins and Port 1998) in the intensity contour

19 Pitch and F0 used here as equivalent terms. For a discussion on these terms see section 2.4

(located as maxima, minima and zero crossings on the derivative contour). This method was chosen for its computational robustness and low computational cost over other automatic vowel/syllable detection methods (see appendix B).

Other features measured (using built-in Praat functions) were jitter, shimmer, harmonics-to-noise ratio and degree of voice breaks. These four features are measures of voice quality (see section 2.4). All of the aforementioned features were also measured on each vowel, in addition to the entire speech segment. The entire process was implemented as a collection of Praat scripts, which can be found in appendix C. Parts of these scripts were included in the development of LinguaTag[20] (Cullen 2008b), a multipurpose speech corpus annotation tool that allows for linguistic transcription, prosodic and emotional annotation of speech and stores the annotation data in XML format for portability.

The extracted feature data was stored in tab-delimited text files that replicate the table-like memory structure used in the scripts. These "table files" can be imported into other programs such as Microsoft Excel® , OpenOffice Calc and MATLAB®. The first two were used for visualization of the data (plots), and the latter was used for the subsequent analysis which is described in the next two chapters.

6.6 Summary

This chapter has described the overall methodology design, as well as some of the "foundation" stages that are shown in Figure 6.2: the design and implementation of the audio recording lab, recording experiments, corpus annotation and feature extraction tools. This work provided the foundation for the analysis of inter-speaker accommodation that is described in the next two chapters. More specifically, chapter 7 describes the analysis and evaluation of prosodic accommodation using the TAMA methodology. Inter-speaker accommodation of temporal features (pauses and overlaps) is discussed in chapter 8.

20 LinguaTag is a product of the SALERO project

7 Accommodation of a/p features

7.1 Overview

This chapter presents the methodology used to validate and describe inter-speaker accommodation of acoustic/prosodic features in human dialogues. The motivation for this work was described in chapter 5. Briefly, a/p features are of interest due to their significant impact on the naturalness of (synthesized) speech in general and SDS in particular. Inter-speaker accommodation of a/p features is therefore a crucial behavioural phenomenon, which SDS could benefit from if it could be adequately described. Thus, the goal is to move away from the "dual monologue" tradition and towards representations that consider the interaction as a whole, rather than the sum of two parts. Such a representation is shown in Figure 6.1 above. The schema hypothesizes the presence of feedback in the interaction. The *Time-Aligned Moving Average* (TAMA) analysis method (Kousidis *et al.* 2008) was designed and implemented to evaluate this hypothesis.

7.2 Design considerations

The initial specifications of the methodology design (see Table 6.1) dictated that the above goal required a continuous, within-dialogue approach. As was discussed in section 4.6, continuous approaches are the only known method of capturing the dynamics of inter-speaker accommodation within a dialogue, which is impossible to do using non-continuous measurements or across-dialogue comparisons. The latter, however have the advantage of being more robust. A trade-off approach is to split the dialogue in two halves, measuring the features on each half and for each speaker, and determining whether the two speakers converged or diverged across the two halves. The major problem of this approach is that there is not enough granularity (or resolution) in the time domain to capture the changing behaviour over time.

Continuous approaches are based on measurements taken on arbitrarily defined time units in the interaction (see section 4.6). In the case of speech, these units can be syllables, utterances, or turns. These linguistically defined units have variable duration. One of the problems encountered is the selection of a single time *point*, which represents the entire unit, for which an average of an a/p feature is calculated. A common solution is that of an arbitrary decision: the start or centre (Nishimura *et al.* 2008) of the utterance have been used, without any mention of the premise for this decision. There is either some underlying assumption that the a/p features are "planned" before the utterance is spoken (or half-way through), or it is simply a matter of convenience. Regardless of the units, the result of this process is a time series of per unit feature values for each speaker. A comparison of the two (or more) time series is used to evaluate the hypothesis of accommodation.

However, a direct comparison is more straightforward if the time points are the same for both series. This is not possible if utterances or other instances of verbal activity are used as units, since two interlocutors do not start and finish speaking at the same time instants. One solution that has been proposed (Nishimura *et al.* 2008; Edlund *et al.* 2009) is to compare actual feature values from one speaker with interpolated values at corresponding time instants from the other speaker. While this solution solves the problem conveniently, it does raise arguments on its validity: unless there is clear evidence that variations of a feature over time can be fitted with a model, then any type of interpolation is unfounded. Fitting a model of feature variation over time is not trivial either, since human speech is not *isochronous*. Another proposed solution is to fit a model on each time series separately and compare the two models. Similarity across the two univariate models is then proposed as evidence of accommodation (Buder and Eriksson 1997; 1999). This approach circumvents the need to "synchronize" measurement points across the two interactants, but also fails to capture the element of feedback: if two series are thought to be inter-dependent, a bi-variate approach is required (Chatfield 1996).

Finally, certain features of human speech (including a/p) are speaker-dependent: pitch (F0), for example, is an inherent property to any individual, as it depends on the size of the larynx - which is why children have higher pitch than adults and female speakers have higher pitch than male speakers. It is also known that speech rate and loudness vary depending on an individual's personality (Oviatt *et al.* 2004). Thus, the manifestation of a/p features in dialogue can be thought of as a combination of inherent traits and dialogue context (accommodation), as schematized in Figure 6.1 (also see section 6.1). In order for a/p feature values of the two speakers to be "compatible", some type of normalization is required. Again using the example above, pitch from a female speaker cannot be directly compared to the pitch of a male speaker. If a direct comparison *is* possible (e.g. for speech rate), then normalization does not have any significant effect.

In conclusion, the points discussed above indicate a bi-variate time series approach, using normalized a/p feature values, which are measured on some kind of synchronous units. This specification led to the formulation of TAMA, which is presented in the next section.

7.3 Time-aligned moving average

The TAMA method utilizes a sequence of contemporaneous fixed-duration frames in which an average value of each a/p feature is calculated. The frames may overlap, making the process similar to a moving average filter, hence the name of the method. The sequence is initiated at the start of the dialogue (time instant zero), and there are two main variables: the *frame length,* and the *time*

step. The frame length is the duration of each frame, while the time step defines the degree of overlap and the total number of frames. The degree of overlap, as a percentage of the frame length is given by the following formula:

$$Overlap = 100 \times \frac{FrameLength - TimeStep}{FrameLength}$$

Equation 7.1: Proportion of frame overlap

The overlap expresses the proportion of a frame that is overlapped by an adjacent frame. Thus a frame length of 20 seconds combined with a time step of 10 seconds yields 50% overlap: the second half of each frame is the first half of the next frame. The total number of frames is given by Equation 7.2 ("\" denotes an integer division):

$$NumberOfFrames = (DialogueDuration \setminus TimeStep) + 1$$

Equation 7.2: Calculation of total number of frames

7.3.1 Frame average calculation

The average a/p feature value of a frame is calculated over the speech intervals found in that frame as shown in Figure 7.1 below. The speech intervals have previously been annotated and a/p features for each interval have been extracted using Praat software (see sections 6.5.2 and 6.5.3).

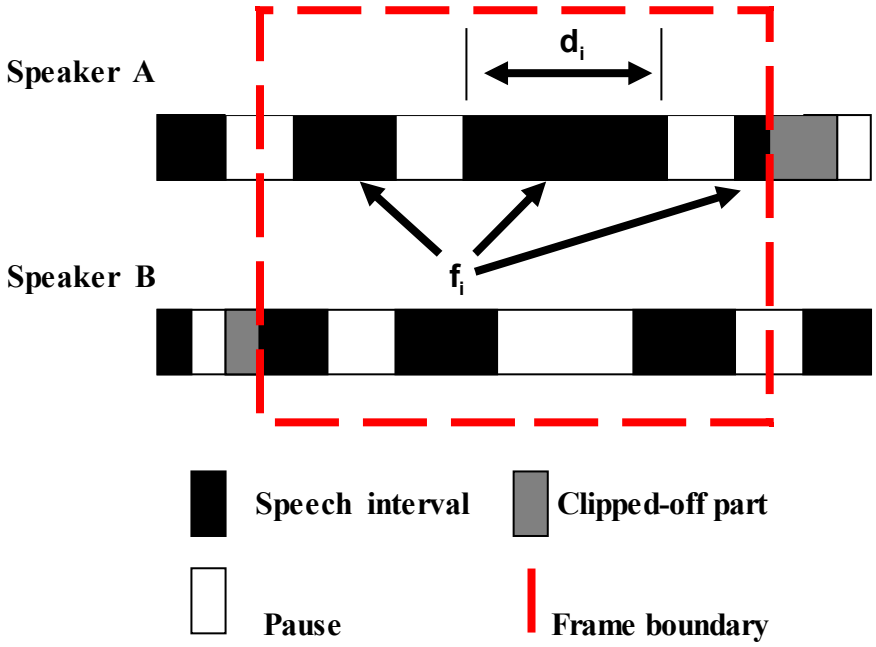


Figure 7.1: Schematic of calculation of TAMA frame average of an a/p feature

Let f_i denote the feature value for speech interval i. The overall mean value of the feature for the entire frame, μ_{frame}, is given as a weighted mean, where the interval durations, d_i are the weights and N is the total number of speech intervals in the frame:

$$\mu_{frame} = \frac{\sum_{i=1}^{N} f_i \cdot d_i}{\sum_{i=1}^{N} d_i}$$

Equation 7.3: Frame average calculation

The weights, d_i can be normalized, if divided by their total, i.e. $w_i = d_i / \Sigma d_d$, with $\Sigma w_i = 1$, in which case the *standard error* is given by:

$$S.E. = \sqrt{\sum_{i=1}^{N} w_i^2 \sigma_i^2}$$

Equation 7.4: standard deviation for weighted mean with normalized weights

where σ_i is the standard deviation of feature f_i in interval i.

The weighting ensures that longer speech intervals have a proportionally higher contribution to the frame average than shorter intervals. The latter are characterized by large variations in their prosodic characteristics: back-channeling expressions often have very low pitch/intensity, while short exclamations have very high pitch/intensity. Since these short intervals are very frequent in spontaneous speech, the averaging would be biased in frames with such intervals. Alternatively, one could concatenate all speech intervals in a given frame and calculate the average feature for the concatenated sound, which leads to the same result: the grand mean of two populations is equal to the mean of the individual means weighted by the population sizes. In this case, the "populations" are the speech intervals, and the "sizes" are the interval durations.

As shown in Figure 7.1, speech intervals may cross frame boundaries. In this case, the duration of the *part* of the speech interval that lies inside the frame is used as the weight in the calculation. This can be thought of as trimming the intervals: the "clipped-off" parts of the speech intervals do not contribute to the frame average. This does not involve a re-calculation of the a/p feature value for the remaining part: the a/p value for the whole interval is used in the calculation and only the duration is affected.

7.3.2 TAMA plots

The result of the process described in the previous section is two series (one for each speaker) of contemporaneous frame averages of a/p features, which can directly undergo bi-variate time series analysis. In order to fully satisfy the specification of section 7.2, the frame averages are normalized by dividing over the overall dialogue mean value, **μ**, of each speaker. This is again calculated using Equation 7.3, considering the entire dialogue as a single frame. An example TAMA plot is shown in Figure 7.2 below.

The TAMA method can be thought of as an expansion of the "half-split" idea (see sections 4.6 and 7.2). Instead of split in two, the dialogue is divided into several shorter frames. The disadvantage in this case, as was mentioned in section 4.6, is that due to the smaller amount of utterances the frame averages tend to be biased by local phenomena, as different utterance types have different prosodic properties. Interrogative statements, for example, have rising intonation, as opposed to declarative statements, which have falling intonation. Thus, there is a trade-off between robustness (longer frames) and resolution (shorter frames). The introduction of overlap, similarly to a moving average filter, has a smoothing effect, highlighting slower-moving (or low-frequency) patterns of prosodic variation over abrupt changes (high-frequency) in prosody that often occur in spontaneous speech.

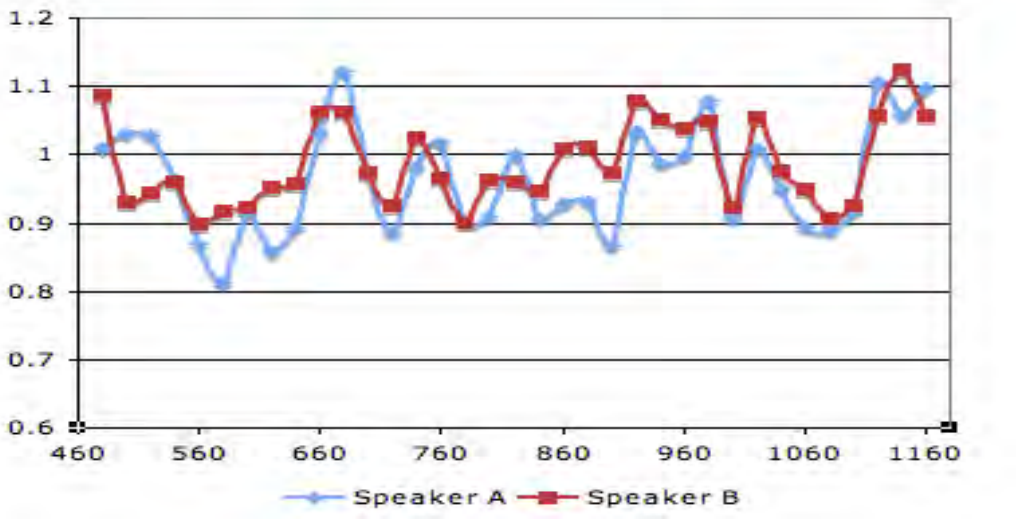


Figure 7.2: Normalized average pitch of two male speakers measured over 30 second frames with 33% overlap (part of dialogue shown)

In addition, the usage of frames, rather than utterances or turns, as units, resolves the issue of synchronous analysis without the need for assumptions over turn allocation to a speaker or marking turn-exchange instants, which is difficult to do in spontaneous speech (Campbell 2009). Instead, a/p feature values are collected by *accumulation* over an arbitrarily defined frame, regardless of the

specific linguistic detail during that time. Some information is lost, such as the time instants that vocalization is initiated or terminated by either speaker. Thus, it is possible that each speaker dominates a different portion of the frame, so that the frame average similarity shown in Figure 7.2 is not indicative of a strictly synchronous similarity in a/p features.

However, speakers in general do not speak contemporaneously most of the time (despite significant occurrences of overlapping speech). In addition, the temporal order of vocalization among speakers is significant when accommodation is considered as a *result* of dialogue structure, rather than an *underlying* behaviour. In naturally occurring human speech, vocalizations can be *anticipated* before they actually occur, thus accommodation does not necessarily depend on the *immediately preceding utterance* o*r turn.* A TAMA frame captures a local portion of the dialogue, and both speakers' contributions during that time are considered as equal in terms of causality. This alleviates the need to define "speaker turns".

Information on each speaker's contribution during a frame is given by Σd_i which, if divided by the frame length, yields a *relative duration:*

$$RelativeDuration = \frac{\sum d_i}{FrameLength}$$

Equation 7.5: Calculation of relative duration

The relative duration has a value between 0 (no contribution) and 1 (entire frame covered by one speech interval of that speaker), and can be used as a confidence score for the a/p value obtained for that frame and speaker: if a speaker's relative duration is low, as a result of minimal contribution, such as a single one syllable back-channeling utterance, it is possible to obtain extremely high or low values for some features. The thresholds depend on the frame length, as longer frame lengths reduce the variance more than shorter frame length. In such cases, points can be removed and replaced by either the overall mean or a linearly interpolated value. Interpolation is justified in this case as each point represents an entire frame rather than a single utterance and thus a linear model *can* be fitted locally for frame averages (if the a/p feature can be assumed to have a normal distribution, see section 7.4.1).

In a preliminary study based on three 30-minute long unconstrained dialogues (Kousidis *et al.* 2008), accommodation was evaluated by visual inspection of the plots for all four a/p features studied (pitch, pitch range, intensity, speech rate). The overall picture was that the two speakers were consistently following each other's prosodic variations over progressively longer time frames (20, 30 and 60 seconds), in all three dialogues. Some dialogue portions, such as the approximately 8

minute-long extract shown in Figure 7.2, showed accurate "tracking" among the two speakers. Several instances of deviation from this behaviour were also found. A careful inspection of these frames showed that the deviation could be attributed to specific causes such as (a) non-standard speech style, such as laughing speech or extreme expressions of enthusiasm (e.g. "wow"), or (b) inaccurate measurements due to low relative duration. While (b) can be dealt with by increasing the frame length, with the consequences discussed in the previous paragraph, (a) is a natural occurrence in human dialogues and it should not be considered as an error. This means that speakers are not *obliged* to converge (accommodate) in their a/p features, rather they do so spontaneously most of the time.

The results in (Kousidis *et al.* 2008) showed that the TAMA method can capture accommodation of a/p features in spoken dialogues, in a continuous representation. In order to formally evaluate this, a statistical validation was sought, as described in the next section.

7.4 Statistical evaluation

As previously mentioned, the statistical method employed to evaluate inter-speaker accommodation was bi-variate time series analysis (Chatfield 1996). This type of analysis considers two time series and is mostly useful when there is indication that the values of one series are in some way dependent on the values of the other series. Time series is perhaps most popular in economics, but has a wide range of applications in such areas as biology, medicine, demographics, as well as engineering (Chatfield 1996).

In the special case of bi-variate time series analysis, one of the two series is considered as the predictor (or independent) variable, while the second series is called the predicted (or dependent) variable. For example, a raise of salaries among a population can be used as the independent variable to predict a raise in household spending. This is a classical example of an *open loop* system: the predicted variable cannot affect the predictor variable in any way. If however both variables are "equal", one of the series is used as the predictor variable by convention. If the predicted variable is found to have an effect on the predictor variable, then *feedback* is present in the process, and the system is called *close loop*. The presence or absence of feedback can be assessed by means of bi-variate time series analysis. In the case of prosodic accommodation, one would expect an open-loop system for uni-directional accommodation (only one of the speakers converges towards the other), or a closed-loop system for bi-directional accommodation (both speakers converge).

7.4.1 Assumptions

Time series analysis considers *stochastic* processes. These processes have the property that they include a random, non-deterministic component. The purpose of the analysis is to de-compose the series into a deterministic and a non-deterministic component. For example, a simple *random walk* is a purely stochastic process: the series starts at zero and at every step it increases or decreases by 1 with equal probability. An *added noise model* is a stochastic process in which each observation is equal to the previous observation plus a random, uncorrelated noise component, ε_i, i.e. $x_i = x_{i-1} + \varepsilon_i$. Stochastic processes describe a wide range of phenomena in which the deterministic component explains some of the variation in the observations, while the variation due to unknown factors is considered as the "random" component. In this study, the time series of a/p frame averages are considered as stochastic processes.

One of the basic assumptions in time series analysis is that of *stationarity.* In its strict form, stationarity requires that the joint probability distribution $F(x_1, \ldots x_N)$ of the observations x_i , of a time series $X = [x_1, x_2, \ldots x_N]$ is *constant over time, i.e.* $F(x_1, \ldots x_N) = F (x_1+\tau, \ldots x_N +\tau)$. However, since this assumption can rarely be satisfied in real applications, the assumption of weak stationarity (or second order stationarity) is more often used (Chatfield 1996). The latter only requires that the mean and variance of the observations x_i need to be constant over time, so that the correlation between two observations of a time series only depends on their time distance, τ. The frame averages can satisfy this assumption, if the true mean and variance of an a/p feature are considered as inherent to the speaker. However, a realization of the process of a/p feature variation during a dialogue does not necessarily exhibit a stationary form. In this case, it is required to transform a series to stationary by using standard techniques such as *differencing* or fitting a model to the series.

A second assumption is that of *ergodicity.* A stochastic process is said to be ergodic if its statistical properties (mean, variance) can be estimated from a single realization of the process: the observed time series is only one possible realization from the probability space comprising all possible realizations of the same underlying process. If the realization is sufficiently long, then the mean and variance of the observed variable can be deduced from this single series of observations. A TAMA a/p feature time series can be considered as a realization of a probability space comprising all dialogues among the two speakers that have the same content (utterance-wise): if the dialogues are sufficiently long, reasonable estimates of the speaker's mean value for a/p features can be obtained..

Conclusively, the underlying assumptions for the individual series of each speaker imply a decomposition of the observed a/p frame averages into a deterministic component (inherent to the

speaker) and a random component which encompasses all other causes of variation: utterance type, mood/emotion, and – most importantly – influence of other speaker. Inclusion of the latter cause of variation as a deterministic component is the purpose of bi-variate analysis.

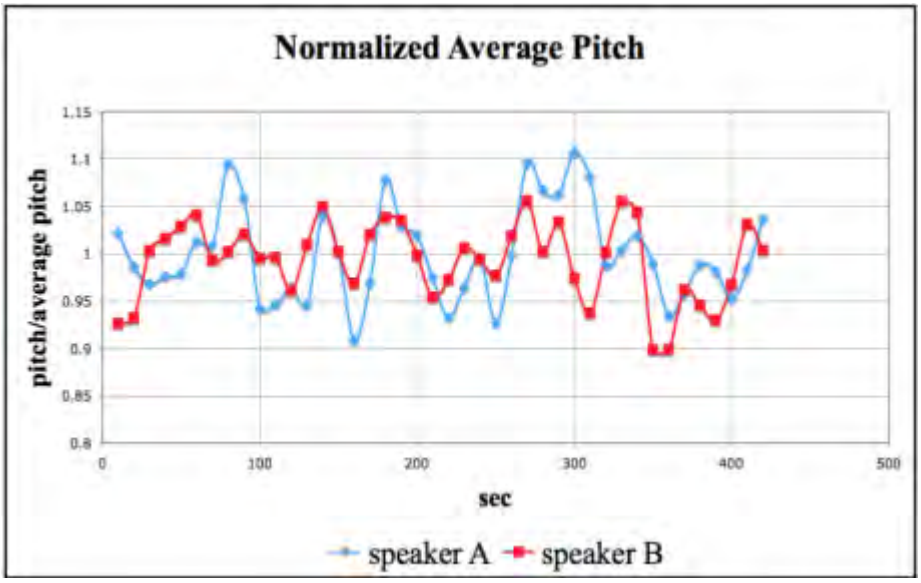
7.4.2 Time series analysis

The first step in time-series analysis is plotting the data, as useful information can be inferred from the time series plots. Two such plots are shown in Figure 7.3. These plots have been obtained from a dialogue recording obtained with the "shipwrecked" experimental scenario (see section 6.4.3). The plots represent the entire duration of the dialogue (approximately 7 minutes). Considering both series in each plot, there is an indication that the two series are correlated, which implies the presence of inter-speaker influence (accommodation). Theoretically, there could be a third, underlying cause that affects both speakers in a similar way, but the only input to the process of dialogue is the two speakers themselves, and no other external factors exist. Thus, the only reasonable conclusion is that the similar movement is the result of influence from at least one of the series on the other.

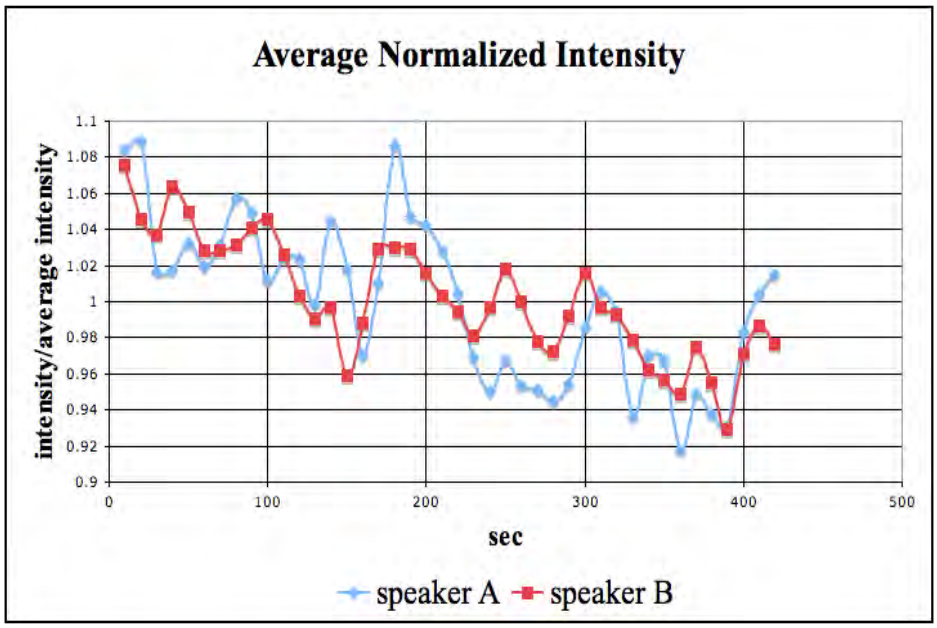
In addition, when considering each series individually, there appears to be a certain degree of *autocorrelation*: consecutive values in the (individual) series are dependent on preceding values of the same series. This is partly a result of the moving average filtering introduced by the TAMA method: each point represents a frame 20 seconds long, 50% of which is shared with the immediately preceding frame. The second underlying cause of autocorrelation is that a speaker's a/p feature average is to an extent dependent on the past values, even if there is no overlap between frames: speakers may well maintain their speaking style over several frames, as is the case of points 3-6 in Figure 7.3a, or exhibit smooth transition from a low to a high value, which indicates that the values are dependent on the preceding values. A final indication of this *autoregressive* structure of the individual series, is that a value above the mean tends to be followed by another value above the mean. The mean in this case is equal to 1, as a result of the normalization method (see section 7.3.1).

Another observation that can be made particularly for the intensity plot (Figure 7.3b) is that the values appear to decline over time. This is an indication of a *global decreasing trend* in the series. A series exhibiting such a trend is not stationary, as the mean value changes over time. A simple method to transform this series to a stationary one is to use *differencing,* i.e. subtracting the preceding value from the current value in the series, creating a new series *Y*, with $y_i = x_i - x_{i-1}$. In some cases, differencing more than once is required. If differencing *d* times is required in order to

achieve stationarity, then the series is called *integrated of order d,* denoted *I(d)*.



(a)



(b)

Figure 7.3: Time series plots of (a) pitch and (b) intensity for two speakers (A,B). Normalized feature averages over 20 second frames with 50% overlap

A useful way of extracting information on individual series, is the *sample autocorrelation function* (ac.f), a good estimate of which is the *correlogram* (Chatfield 1996). A correlogram is a plot of correlation coefficients over a number of *lags.* A lag of 1 denotes the immediately preceding value of the series, a lag of 2 denotes the value before that, etc. The sample autocorrelation coefficient, r_k at lag k is given by:

$$r_k = \frac{\sum_{t=k+1}^{N} (x_t - \mu)(x_{t-k} - \mu)}{\sum_{t=1}^{N} (x_t - \mu)^2}$$

Equation 7.6: Sample autocorrelation coefficient

where μ is the overall mean for the entire dialogue, x_t is the TAMA frame a/p feature average for frame t. both calculated by Equation 7.3, and N is the total number of TAMA frames.

The correlograms of the first six lags for the two series in Figure 7.3a are shown in Figure 7.4. The horizontal bars denote confidence intervals at $\pm 2/\sqrt{N}$.

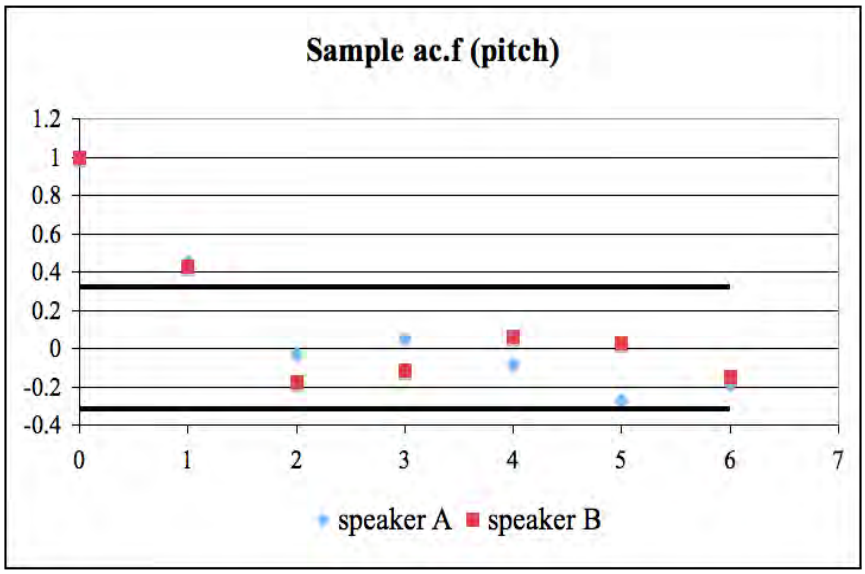


Figure 7.4: Correlograms of the two individual series shown in Figure 7.3a

The coefficients for both series quickly drop to zero, which indicates that the series are stationary (values within the confidence intervals are statistically zero). The coefficient at lag zero is always equal to 1 (series correlated with itself). There is one significant coefficient at lag 1, with a value around 0.4 for both series. This validates the hypothesis of the autoregressive structure of the series as the values for each series are dependent on the immediately preceding values.

In contrast, the correlograms for the two intensity series (Figure 7.3b) are typical of series exhibiting a global trend (see Figure 7.5): the coefficients decline exponentially, but remain significant and do not drop to zero. Thus, a transformation (such as differencing) of these series is required in order to obtain two stationary series.

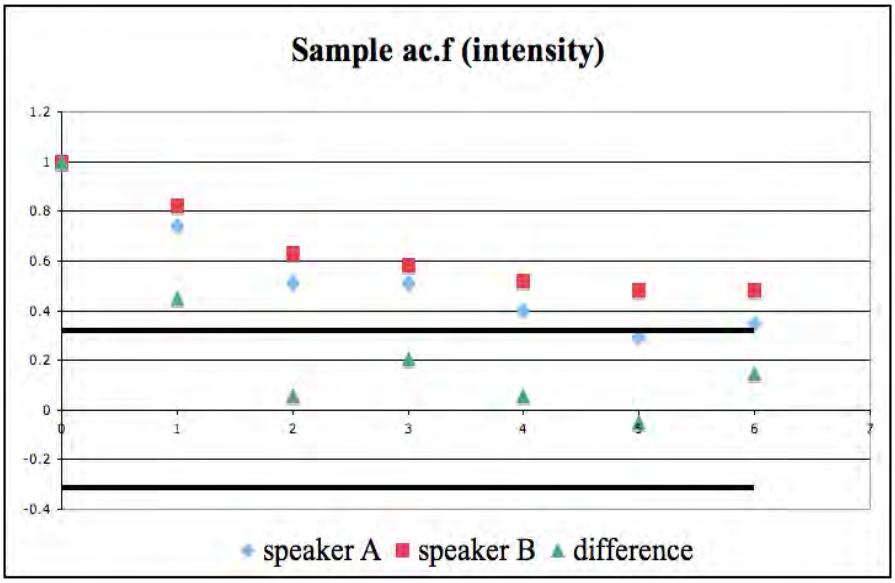


Figure 7.5: Correlograms of the two individual series shown in Figure 7.3b

The analysis so far has concentrated on each individual series. As mentioned earlier, a bi-variate analysis is required to validate the hypothesis of feedback between the two series.

7.4.3 Bi-variate analysis

Bi-variate analysis considers two series together and is the simplest case of *multivariate* time series analysis. Each observation, rather than being a real number, is a *vector*, the elements of which are the values from each individual series. If x_1, x_2 are two time series, then the vector $X = [x_1 \; x_2]$ is the bi-variate time series. In general, the observations of an n-variate time series are $n \times 1$ vectors. The individual univariate time series are called *component series* (Chatfield 1996).

The relationship between the component series can be explored by means of the *sample cross-correlation function* (cc.f), an estimate of which is the cross-correlogram. In order to to obtain a cross-correlogram, one needs to distinguish the component series into an input, x (independent variable), and output, y (dependent variable). As mentioned in section 7.4, this is done arbitrarily in this case, as the two subjects have an equal role in the dialogue experiment of the shipwrecked scenario (see section 6.4.3). In this manner, the series for speaker A is considered as the "input", and the series of speaker B is considered as the "output".

Careful consideration needs to be given to cross-correlation, as spuriously large coefficients may appear in the cross-correlogram if the component series are themselves autocorrelated (Chatfield 1996). A technique commonly used in such cases is that of *pre-whitening* the component series. This means that their correlograms should resemble *white noise*, which is a random process in

which subsequent values are uncorrelated[21]. Therefore, each component series has to be transformed so that its respective correlogram shows no significant coefficient. In the case of the two pitch series (Figure 7.3a), this can be achieved by fitting an autoregressive (AR) model of order 1 to each series. This is indicated by the respective correlograms of the series (Figure 7.4), which show a significant coefficient at lag 1 for both series. According to (Chatfield 1996), the value of that coefficient is the best estimate for an *alpha* (**α**) value in an AR(1) model of the form $(x_i - \mu) = \alpha(x_{i-1} - \mu) + \varepsilon_i$, where ε_i denotes random noise. Using the value of $\alpha = 0.4$ found on the correlogram, the above equation is solved for ε_i, which yields a *residual series* for each speaker. The success of the pre-whitening method can be validated by plotting the correlograms of the residual series, in order to determine whether any coefficients remain significant (not shown).

Cross-correlation coefficients are then calculated for this pair of residual series. The sample correlation coefficient r_k at lag k is given by:

$$r_{xy}(k) = \begin{cases} \dfrac{\sum_{t=1}^{N-k} (x_t - \mu_x)(y_{t+k} - \mu_y)}{\sqrt{\sum_{t=1}^{N} (x_t - \mu_x)^2 (y_t - \mu_y)^2}}, & k \geqslant 0 \\ \dfrac{\sum_{t=1-k}^{N} (x_t - \mu_x)(y_{t+k} - \mu_y)}{\sqrt{\sum_{t=1}^{N} (x_t - \mu_x)^2 (y_t - \mu_y)^2}}, & k < 0 \end{cases}$$

Equation 7.7: Sample cross-correlation coefficient

where μ_x, μ_y are the means of the component (residual) series x,y respectively, x_t, y_t are the values of the residual series at time t, and N is the total number of points in the residual series. The cross-correlogram for the two pitch series (Figure 7.3a) is shown in Figure 7.6 below.

One major difference between the cross-correlogram and correlogram plots is that the former contains both positive and negative lags. According to (Chatfield 1996), a linear system with input x and output y demonstrates *feedback* if significant coefficients are found at zero or positive lags. However, if the roles of the two speakers' series – as input and output – are reversed, then the coefficient at lag 1 which can be seen in Figure 7.6 will appear at lag -1. Therefore, a coefficient at lag 1 or -1 is an indication of uni-directional convergence, in this case A→B: as the roles are

21 For a formal definition of white noise, see Chatfield (1996)

reversed, B is now the input and A the output, and a significant coefficient at lag -1 means that A converges to B. This can be seen on several occasions in Figure 7.7, where A (blue) is lagging behind B (orange) by one point, particularly in the right part of the plot.

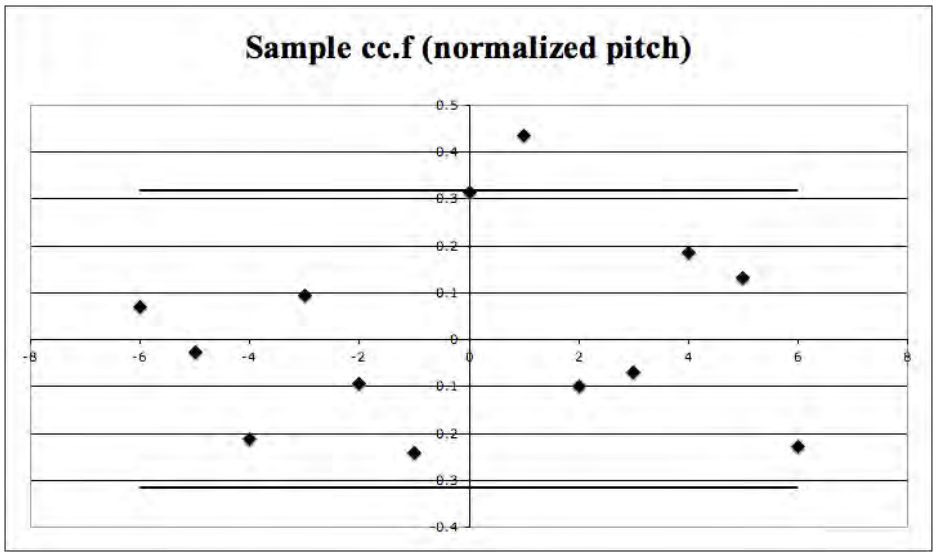


Figure 7.6: Sample cross-correlogram of the two series of Figure 7.3a, pre-whitened by fitting an AR model with α = 0.4

It is noted that this interpretation of the cross-correlogram is not very reliable, due to the presence of a (borderline) significant coefficient at lag zero. This indicates the presence of feedback in the system, unless a common underlying process is affecting both series. This point is emphasized because correlation by itself does not imply causality: unless the possibility of a common external factor can be safely excluded, there is no basis to assume a causal relationship. Since the only input in the dialogue is provided by the speakers themselves, the coefficient at lag zero has to be attributed to feedback (see section 7.5). When feedback is present, the interpretation of the correlogram can be misleading (Chatfield 1996), especially in terms of using the cross-correlogram in order to estimate model parameters, e.g. as in the univariate case, where it was possible to estimate the alpha value for an AR(1) model directly from the correlogram.

In Figure 7.7, the residual (pre-whitened) series are plotted. These residuals represent the amount of variation in the a/p features *not accounted for by autocorrelation* (a deterministic component). The existence of one or more significant cross-correlation coefficients implies the existence of an additional deterministic component, whether an external factor that affects both series, or a *causal relationship* between the two series (the latter in this case). However, estimation of the power of this component is not possible using the correlogram *because* of feedback: as shown in Figure 7.7, there are points at which the two series are "in-phase", as well as points at which blue is lagging behind

orange. These two coefficients are competitive: the instances of zero lag reduce the value of the coefficient at lag 1 and vice versa. Positive and negative lag coefficients are also competitive. In fact, in an extreme case where two pure open-loop processes with opposite lags (at -1 and 1) are combined (concatenated), there is only one significant coefficient at lag 0. Therefore, the values of the cross-correlation coefficients can only be used for model parameter estimation only if it is certain that there is *no* feedback.

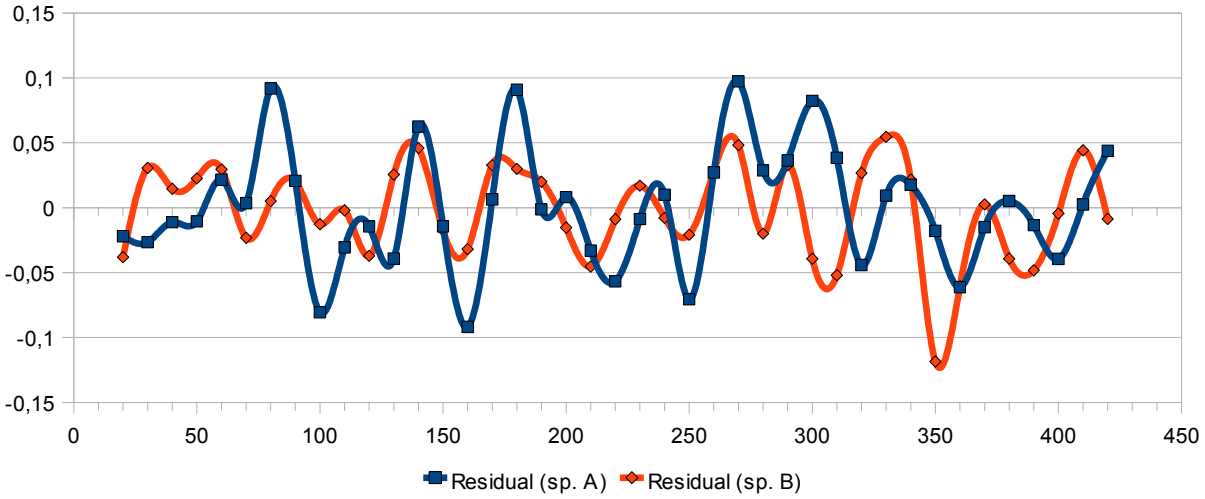


Figure 7.7: Residual series plot for the two series in Figure 7.3a after fitting an AR(1) model with $\alpha = 0.4$ to both series

In addition, each point in the time series represents an entire frame, rather than a single time instant; therefore, the coefficients at lag 0 and and lag 1 are competitive with respect to the frame length. In other words, some of the autoregressive structure is "masked" due to the averaging process. Intuitively, accommodation in human dialogues is always deterministic, as speakers accommodate to each other's speech based on *past* utterances. However, it has been suggested (Heylen 2009) that feedback in human interaction can be *instantaneous,* due to visual or other cues. In the absence of visual feedback in the recordings analyzed here, it can be argued that instantaneous feedback occurs by means of overlapping speech segments. As pointed out in section 7.4, feedback implies bi-directional accommodation (A↔B). However, due to the issues discussed here, i.e. the competitiveness between coefficients and the loss of some temporal information due to the frame length, the cross-correlogram cannot show the degree of convergence *separately* for each speaker. Despite the fact the cross-correlogram is not useful for model *estimation,* it can be used for model identification (see section 7.4.4).

In a paper presenting this statistical evaluation method (Kousidis *et al.* 2009a), five dialogues from the "shipwrecked" scenario corpus were analyzed for accommodation of four a/p features: pitch,

intensity, pitch range and speech rate (see Table 7.1). Significant positive correlations were found for all four features, albeit mostly for pitch and intensity. Most of these coefficients were found at lag zero, which implies bi-directional accommodation. Whether uni-directional or bi-directional, the presence of a significant positive correlation coefficient constitutes a statistical validation of accommodation, as there is a deterministic component for at least one of the speakers that is caused by inter-speaker influence.

Dialog	Significant coefficients (lags)			
	Pitch	Intensity	Pitch range	Speech rate
1	0,1	0,1	1	-1
2	0	0	0	-
3	1	0	-	1
4	0	0	0	-
5	0	0	-	0

Table 7.1: Lags at which significant positive cross-correlation coefficients are found among two speakers in 5 "shipwrecked" dialogue recordings

Importantly, the positive sign of the cross-correlation signifies *convergence,* in other words adaptation of one's a/p features to the respective features of the other. This occurs simultaneously along different dimensions (or modalities), if each a/p feature is though of as a distinct channel of accommodation. A negative cross-correlation coefficient would signify *divergence,* or non-accommodation (see section 3.4.3), but no negative coefficients were found in (Kousidis *et al.* 2009a). As positive and negative coefficients are also competitive at the same lag, non-accommodation will not be statistically significant unless it occurs in a relatively large portion of the dialogue. The results of (Kousidis *et al.* 2009a) were confirmed from the analysis of the rest of the corpus (see appendix A).

7.4.4 Modeling inter-speaker accommodation

A major motivation for describing inter-speaker accommodation phenomena, apart from gaining a better understanding of the phenomena, is to provide a model that can be used in SDS implementations. As was discussed in section 5.1, such an implementation is desirable for improving on the naturalness as well as the efficiency of SDS. This section describes possible modeling approaches.

The presence of autocorrelation and feedback in the bi-variate process of accommodation points towards a vector autoregressive (VAR) model. This is the multivariate extension of the AR model

that was used in section 7.4.3 in order to "pre-whiten" the two component series. The simplest possible model is the VAR(1), a model which takes into account the preceding values of the series in order to predict (or forecast) the current values:

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \Phi \begin{bmatrix} x_{i-1} \\ y_{i-1} \end{bmatrix} + E \,, \Phi = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix}, E = \begin{bmatrix} e_x \\ e_y \end{bmatrix}$$

Equation 7.8: A VAR(1) model

where *x,y* are the two component series, Φ is the *parameter matrix*, and *E* is the *error vector*. The elements φ_{ij} of the main diagonal (φ_{11} ,φ_{22}) in the parameter matrix are the *autoregressive* terms, which explain the autoregressive portion for each series (the AR models). The secondary diagonal elements (φ_{12} ,φ_{21}) are the *feedback* terms (Chatfield 1996): If both are significantly large, the system is closed-loop and demonstrates feedback. If the matrix Φ is triangular, i.e. one of the feedback terms is zero, then the system is open-loop, which implies unidirectional accommodation. If both feedback terms are zero, then there is no cross-correlation and Equation 7.8 yields two separate univariate models.

Estimation of the parameters can be performed by means of error minimization. Let *x,y* be two series of TAMA frame averages (mean intensity in dB), as shown in Figure 7.8:

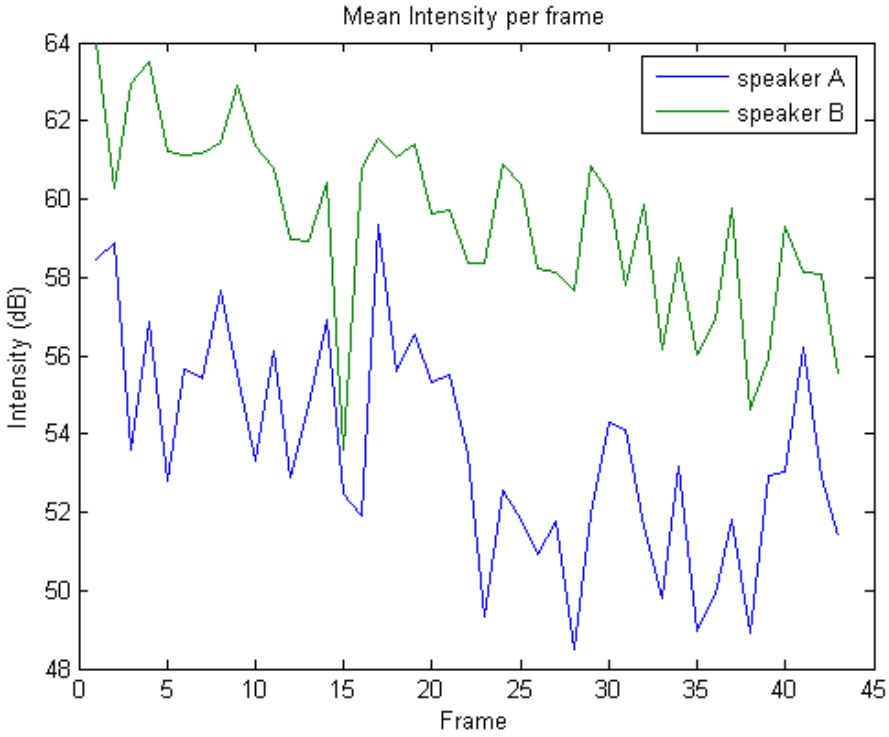


Figure 7.8: TAMA plot of average intensity for two speakers A,B

Both series exhibit a global decreasing trend as shown in the correlograms of Figure 7.9a (the coefficients do not decline to zero). Unfortunately, differencing (see section 7.4.2) in this case results in a large negative autocorrelation coefficient at lag 1 in the correlogram for both series (see Figure 7.9b). This is a sign that the series have been *over-differenced.* Instead of differencing, an AR(1) model is fitted for both series with the following method (the correlogram does not provide an estimate for an alpha value in this case, as the series are not stationary yet):

Using the AR(1) model equation $(x_i - \mu) = \alpha(x_{i-1} - \mu) + \varepsilon_i$, a *least squares fit* is performed in order to obtain the *slope* (the offset is ignored). Thus, if $y = (x_i - \mu)$ and x is the *lagged* series of y: $(x_i = y_{i-1})$, then solving the least squares problem of the form $y=\alpha x+\beta$ yields an alpha value of 0.42 for speaker A and 0,40 for speaker B. The *residual series* are now stationary (see Figure 7.9c).

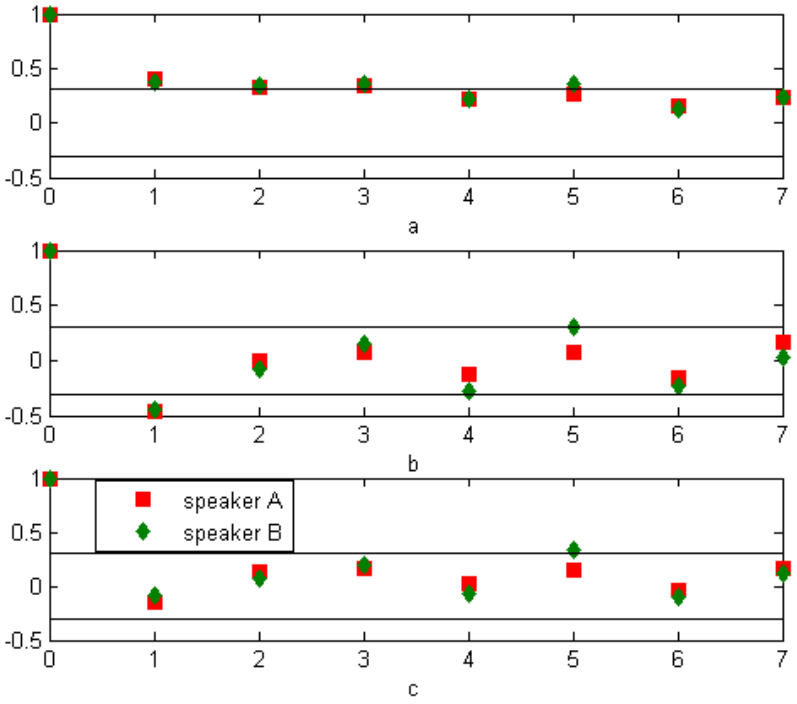


Figure 7.9: Correlograms for the two series of figure 7.8: (a) original series, (b) differenced series and (c) series fitter with AR(1) models

Now the cross-correlogram can be calculated. A significant correlation is found for lag zero, as expected (not shown). In order to compute the *feedback* terms, the same process as in the univariate case is followed by performing *multiple* regression. Equation 7.8 can be written as a set of simultaneous equations:

$$\hat{x}_i = \varphi_{11} x_{i-1} + \varphi_{12} y_{i-1} + \varepsilon_x$$
$$\hat{y}_i = \varphi_{21} x_{i-1} + \varphi_{22} y_{i-1} + \varepsilon_y$$

Equation 7.9: VAR(1) model written in simultaneous equation form

Solving the multiple least squares problem for $y=\alpha_1\chi_1 + \alpha_2\chi_2 + \alpha_0$ yields:

$$\Phi = \begin{vmatrix} 0.18 & 0.44 \\ 0.08 & 0.34 \end{vmatrix}$$

The feedback term for speaker A is large (0.44), which implies that speaker A *converges* towards speaker B. The feedback term for speaker B is insignificant, which implies no convergence from B. However, the above model does not take into account the lag zero correlation between the two speakers. As discussed above, the lag zero coefficient accounts for the accommodation taking place within the TAMA frame time-span, which includes instantaneous feedback. A third deterministic component can be added to the model in order to account for lag zero cross-correlation:

$$\hat{x}_i = \varphi_{11} x_{i-1} + \varphi_{12} x_{y-1} + \theta_1 y_i + \varepsilon_x$$
$$\hat{y}_i = \varphi_{21} x_{i-1} + \varphi_{22} x_{y-1} + \theta_2 y_i + \varepsilon_y$$

Equation 7.10: VAR(1) model with added zero lag component

where θ_1,θ_2 are the *zero lag feedback terms.* Multiple linear regression yields:

$$\Phi = \begin{vmatrix} 0.15 & 0.24 \\ 0 & 0.15 \end{vmatrix}, \Theta = \begin{vmatrix} 0.56 \\ 0.46 \end{vmatrix}$$

Large zero-lag feedback terms are found for both speakers. In fact, it is apparent most of the accommodation occurs within the 20-second long TAMA frame, although at least one of the speakers (A) is accommodating based on even older context (see Figure 7.10).

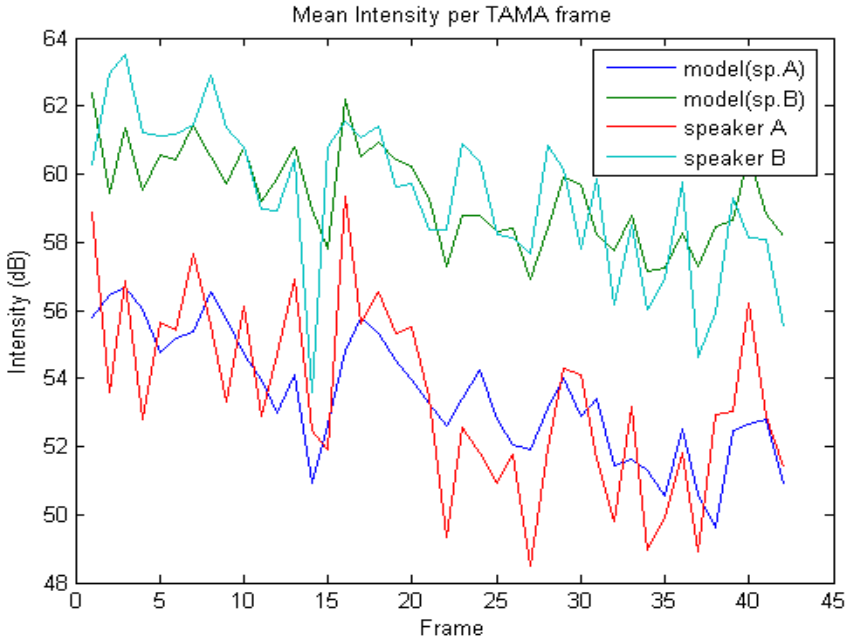


Figure 7.10: TAMA frame series (mean intensity) fitted with VAR(1) model with zero-lag feedback term

Another approach is to bias the fit in favor of autocorrelation, i.e. to *fix* the autoregressive terms to the values calculated for the univariate case. The purpose of this is to enforce the hypothesis of no accommodation, as was done in the calculation of the cross-correlogram. In this case, multiple regression is performed for the *residual series* occurring *after* the optimal AR model has been fitted to each series, i.e:

$$\hat{\varepsilon}_{x,i} = \varphi_{12} y_{i-1} + \theta_1 y_i$$
$$\hat{\varepsilon}_{y,i} = \varphi_{22} x_{i-1} + \theta_2 x_i$$

Equation 7.11: VAR model with feedback terms at lags 0 and -1 fitted to residual series

where $\varepsilon_x = x_i - \alpha x_{i-1}$ is the residual series of the AR(1) model with the optimal α (0,42 and 0,4 for speakers A, B respectively). The least squares fit yields:

$$\Phi = \begin{vmatrix} 0.42 & 0.05 \\ -0.11 & 0.40 \end{vmatrix}, \Theta = \begin{vmatrix} 0.54 \\ 0.40 \end{vmatrix}$$

Therefore even when " fixing" the autoregressive terms, the lag-zero feedback terms remain large. However, the lagged feedback terms have become insignificant for both speakers in this model (see Figure 7.11)

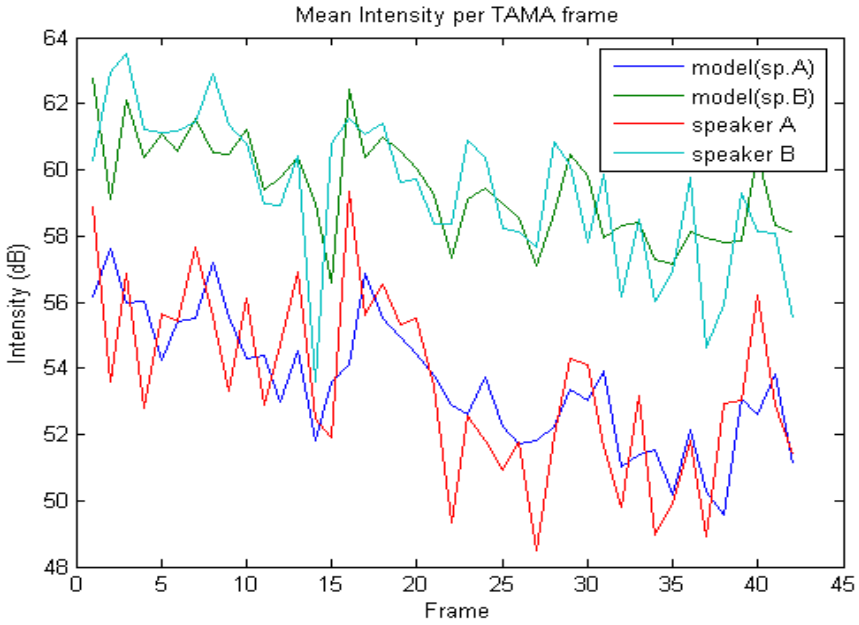


Figure 7.11: TAMA frame series (mean intensity) fitted with AR(1) models and VAR(1) model with zero-lag feedback terms on residual series

As shown in figures 7.10 and 7.11, both models follow the local variations of the speakers' intensity adequately. For comparison, the mean square error (MSE) for the models are shown in Table 7.2.

The models should not be expected to follow the actual values *accurately,* because variations in a/p features such as intensity are not subject solely to inter-speaker accommodation. Other factors that influence the a/p features (such as the utterance type or phonetic content) can be added to the model as *exogenous factors,* resulting in a VARX model (Chatfield 1996). These are added to the right-hand side of Equation 7.10 and constitute additional deterministic components to the process. For example, a promising extension of the TAMA methodology would be to annotate the dialogues for utterance type (declaration, wh-question, yes/no question, back-channel, etc.) and calculate an average value per feature and per utterance type for the whole dialogue. The contribution of each utterance to the frame average would then be based on its *normalized* value, i.e. its value relative to the *utterance-type mean.*

speaker	VAR(1) with zero lag term	VAR(1) with zero lag term and fixed autoregressive terms
A	**3,835**	**4,165**
B	**3,093**	**3,272**

Table 7.2: Mean square error (MSE) for the model in figures 7.10 and 7.11

Another possible refinement of the modeling method described here would be to consider *co-integration.* In Figure 7.5, the correlogram of the two intensity series shows that they are not stationary (the coefficients do not decline to zero). However, their first order difference *is* stationary. This means that the two series are co-integrated, and the order of co-integration is equal to 1 (Chatfield 1996). It is possible then to simplify an otherwise complicated model by including the *co-integration* vector **[1 -1]** in the model. This approach could be given meaning by positing that accommodation may (partly) be affected by the *distance* (or perceived distance) between the speakers along a hypothetical continuum of accommodation/non-accommodation.

7.5 Discussion

This chapter has presented a novel methodology (TAMA) for describing accommodation of a/p features in spontaneous dialogues. The main advantages of the methodology are (a) the continuous representation of accommodating behaviour, (b) the acquisition of two time series which can be statistically analyzed to validate the hypothesis of accommodation, (c) the robustness of the frame average estimation by means of overlapping frames, and (d) feature independence, provided that the feature has a measurable magnitude and sufficient amount of data is included in the frame. In addition, the hypothesis of accommodation was statistically verified by means of bi-variate time series analysis, and the direction and degree of accommodation were quantified by means of

statistical modeling of the VAR variety.

Accommodation of a/p features has been previously observed and statistically evaluated, by comparing the a/p feature averages of entire dialogues (Oviatt *et al.* 2004; Suzuki and Katagiri 2005). The latter studies addressed the issue of describing inter-speaker accommodation within a dialogue by splitting the dialogue in half and comparing a/p feature averages across the two halves. TAMA builds upon the idea of "half-split" dialogues, but extends it to any number of dialogue parts, which are termed dialogue frames. This leads to a combination of merits from utterance-based continuous representations and across-dialogue comparisons. The trade-off between resolution and robustness is addressed by allowing frames to overlap. Thus, TAMA yields a continuous representation of accommodation phenomena in the form of two time series.

Existing work on describing a/p accommodation by means of time series differs from TAMA in various key points, but there are also significant similarities. (McRoberts and Best 1997) used the same normalization method as TAMA (dividing an F0 measurement over the speaker's overall average F0) and presented time series plots of F0 variation. However, the measurements in that study were taken on each utterance. TAMA avoids attributing turns to each speaker, which is difficult to do in spontaneous speech, and is more consistent with representations that consider dialogue as a synchronous activity (Campbell 2009; Heylen 2009). This point is further elaborated by the statistical analysis which reveals a significant lag-zero term for both speakers in the dialogues studied.

(Kakita 1996) also used a time series approach in order to study accommodation of F0, but used scripted dialogue rather than spontaneous speech, and measured F0 on a specific syllable in a word that was present in each utterance by design. In addition the F0 values were not normalized, and thus inherent F0 of speakers was not taken into account. (Buder and Eriksson 1997; 1999) used a time series approach to compare synchrony of F0 and Intensity "cycles" across two speakers over floor exchanges. The sinusoidal models were fitted on each series separately, and thus the analysis was not bi-variate and could not reveal the presence of feedback.

Perhaps the most similar approach to TAMA is that of (Nishimura *et al.* 2008), which used a lag-zero cross-correlation to assess accommodation of F0, and also calculated a bi-variate model for continuous system adaptation of F0 towards that of the user. However, (Nishimura *et al.* 2008) used utterances as units and analyzed small (minute-long) portions of dialogues. The findings were similar to those presented here and in (Kousidis *et al.* 2009a): Significant lag-zero correlation of F0 and a model that has to take instantaneous feedback into account.

The statistical approach (bi-variate time series analysis) presented in section 7.4.3 bears

resemblances to that of (Jaffe *et al.* 2001). The latter study, which focused on rhythmic features (see section 4.5.6), used frame lengths of 5 seconds in which the average duration of vocal states was measured. (Jaffe *et al.* 2001) also accounted for auto-correlation by fitting AR(2) models to the individual series prior to performing lag regression of frame averages, i.e. the regression strength (R^2) between each series and the lagged series of the interlocutor was calculated (for up to 12 lags). This regression strength was interpreted as an indication of coordination among the two speakers, as well as of the strength and direction of accommodation. Importantly, (Jaffe *et al.* 2001) excluded lag zero from the analysis. In contrast, (Kousidis *et al.* 2009a) used the cross-correlogram as an indication of accommodation, and the feedback terms of the VAR models as indicators of the strength and direction. As the VAR models were also calculated by linear regression (least squares), the feedback terms can be interpreted as the *slope* of the fitted line.

The statistical analysis (section 7.4.3) revealed significant cross-correlation at lag zero and/or neighbouring lags (less often), which was considered as indicative of feedback, the physical interpretation of which is bi-directional accommodation. This interpretation, schematized in Figure 6.1, is only valid if the possibility of any external factors influencing the prosodic features of the two speakers can be excluded, as correlation by itself does not imply causality (Chatfield 1996). However, a/p features carry several functions, as discussed in section 2.4.2. The possibility that one of these functions is influencing both speakers simultaneously, leading to a significant coefficient at lag zero has to be considered thoroughly.

Any linguistic functions of prosody have to be excluded, because that would imply that speakers produce utterances which have the same or very similar lexical, semantic and pragmatic content. That would only occur if speakers were repeating each other's utterances. Of course, there is the frequent phenomenon of users complementing each other's utterances, thus adhering to the original utterance intonation structure. However, such behaviour would have to be characterized as accommodation.

Paralinguistic functions are less trivial to discard. The frequency code (Gussenhoven 2005), for example, carries the function of dominance (see section 2.4.2). Therefore a simultaneous rise/fall in average F0 could be interpreted as a dominance “duel” between the two speakers. However, there is no indication of such behaviour in the recorded dialogues, in which speakers are eagerly cooperating and generally enjoying the sessions, as the frequent occurrences of laughter suggest. The effort and production codes are mostly manifested in local pitch and intensity variations which would contribute little to a 20 second frame average.

Emotional content may also influence the a/p features of the speakers. Considering a dimensional

approach (Schroeder 2004) which would be more appropriate for these recordings than full-blown emotional categories, all four features are positively correlated to *activation.* Therefore, simultaneous activation, which is likely to occur as a result of stimuli arising from the progress in the task, as in the case of the "shipwrecked" recordings with a ranking score (see section 6.4.3), would result in simultaneous rises in the a/p features. However, such stimuli are a few distinct events in the dialogue, but the TAMA plots reveal synchronous variation throughout the dialogue. If simultaneous activation occurs as a result of stimuli introduced by the speakers themselves, then that behaviour would have to be characterized as accommodation, as in co-activation manifested by similar prosodic variations.

Therefore, all *known* functions of prosody can be excluded as *external* causes of simultaneous prosodic variations measured by means of the TAMA methodology. The only input to the process of dialogue are the utterances of the two speakers: contemporaneous activation exhibited by similar prosodic manifestation is therefore a *result* of the interaction, rather than a cause. Therefore, the correlation can be attributed to inter-speaker accommodation. The ubiquitous nature of the phenomenon points to an implementation in an SDS environment based on the models derived in section 7.4.4. Such an implementation is presented in chapter 9.

Another important point relates to the deterministic nature of accommodation, i.e. the accommodation of speakers to their partners' past utterances. This is schematized in Figure 6.1 as a continuous feedback loop which follows a deterministic circular path from speaker A to speaker B and vice versa. This type of description implies a succession of turns between the two speakers. However, spontaneous dialogues are characterized by overlapping speech and "fuzzy" turn successions (this is further discussed in the next chapter). Therefore, the overlapping and otherwise perplexed speech segments point to instantaneous accommodation, as schematized in Figure 2.3 (Heylen 2009). This is captured by the lag zero coefficient and the zero-lag feedback terms of the models, although these measures also express some of the autoregressive accommodation due to the fact that each point in the series represents a time *span* rather than a time *instant.*

Intuitively, it could be argued that accommodation is always deterministic, as a/p features of overlap segments accommodate to the spoken part of the utterance being overlapped. However, the purpose of the overlapping segment could be to signal understanding and prompt the speaker to proceed with their point at speed (indicated by the a/p features of the overlap segment). The speaker would then accommodate to this stimulus while the overlap segment is still being vocalized (or even before if the overlap can be predicted by the speaker). This type of behaviour is a prime example of instantaneous feedback. In fact, this type of feedback is essential to the organization of the dialogue,

and is manifested in several modalities. In the absence of visual contact, this function is carried by overlap segments. This also explains the findings of (Bosch *et al.* 2005), in which significantly more overlaps were found in telephone conversations in comparison to face-to-face conversations.

8 Accommodation of temporal features

8.1 Overview

This chapter presents work undertaken in order to describe inter-speaker accommodation of timing in human dialogues. The motivation for this work was outlined in chapter 5. Briefly, the time instants at which speakers initiate/terminate their vocalizations during dialogue are of interest because they characterize the *floor* transitions between them. Therefore, it would be desirable to describe this process, in order to implement more natural (and by extension more efficient) interaction management strategies for SDS, which are currently mostly based on a "ping-pong" or "half-duplex" model: the human user and the automated talking agent are speaking in *turns,* where a turn is defined as a time interval during which one of the parties holds the floor (speaker), while the other party is concentrated on understanding and processing information (listener). When the turn is exchanged, the parties switch roles and the process is repeated in the opposite direction. Turns can be exchanged either when the floor is released by the speaker (inter-speaker silence), or if the listener interrupts the speaker in order to "take over" the floor (overlapping speech).

However, the above account is clearly insufficient in describing natural human speech. The latter is characterized by frequent instances of overlapping speech which cannot be characterized as turn exchanges. It is widely accepted that one of the main functions of these *overlaps* is to provide *feedback* to the speakers that is currently holding the floor, and that this feedback is essential in the process of dialogue. In its absence, the "speaker" cannot assess whether the "listener" has understood what is being said, which makes the communication inefficient. Feedback is not necessarily verbal, but can occur on other modalities, such as head nods or eye movement. Therefore, when these modalities are not available (e.g. in telephone conversations) a greater amount of verbal feedback is expected. Yet a "half-duplex-plus-feedback" model is still insufficient in characterizing human interaction, and attributing a specific communicative function (e.g. "feedback" or "declaration") to each utterance in a dialogue is not without problems (Bosch *et al.* 2004b). As a result, attributing turns to each speaker is not a straightforward task and requires a certain degree of subjectivity in order to be achieved (Raux 2008), as was discussed in section 2.3.2.

Categorization of dialogue acts into different types and segmentation of the dialogue into semantically and pragmatically defined turns is the subject of conversation and discourse analysis (see section 2.4.4). Within this field, the temporal structure of human interaction is considered as accommodating the smooth transition of turns between the two speakers. In contrast, research on inter-speaker accommodation focuses on adaptation of the speakers' temporal patterns in order to match each other's "temporal behaviour". These two approaches are distinct in origin but involved with the same phenomena, namely the temporal structure of dialogue. The same is true for research

in other modalities, such as lexical and syntactic choice (Matessa 2001): both conversation analysis and accommodation theory explain the tendency of speakers to make similar lexical and syntactic choices. The two lines of research are complementary and one can benefit from the other.

In this context, this chapter presents an application of the TAMA methodology to temporal features as well as a novel dialogue representation, in order to describe temporal accommodation phenomena. The former lies exclusively within the accommodation theory line of research, while the latter "crosses over" into conversation analysis by considering the influence of *turn share distribution* (a measure of dialogue activity and speaker dominance) on the same temporal features. This cross-over is unavoidable, as pauses and overlaps are the main features in any description of the temporal organization of dialogue.

8.2 TAMA analysis of temporal features

The first approach towards describing inter-speaker accommodation of temporal features comprised an adaptation of the TAMA methodology described in chapter 7. As was discussed in section 7.5, TAMA is a feature-independent method for describing accommodation phenomena, provided that calculating an average value is feasible as well as meaningful. Two such features were identified from the literature review (see section 5.2) that are related to the temporal organization of dialogue: the duration of silent intervals and the occurrence of overlaps.

Importantly, the silent intervals (or pauses) occur both between speech intervals of the *same speaker* as well as during *floor exchanges*, i.e. a different speaker resumes speaking after the pause. (Edlund *et al.* 2009) used the terms *gap* and *pause* to differentiate between these two conditions, respectively. Here, the term *switch pause* is used when a different speaker vocalizes after the pause, while *pause* is used to signify *either or both* conditions, depending on the context or explicitly disambiguated. This is because accommodation of pause length has been known to occur mainly for switch pauses (but see Jaffe *et al.* 2001), but also because in the proposed representation the two are not differentiated (see section 8.3.2).

Similarly, overlaps can be interrupting and non-interrupting. Interrupting overlaps occur when the speaker that "barges-in" takes over the floor after the overlapping segment, while non-interrupting overlaps result in the original speaker retaining the floor. In this text, these two cases are simply termed interrupting and non-interrupting overlaps. Typical non-interrupting overlaps comprise back-channel feedback utterances which are not necessarily proper phrases or words in a strict linguistic sense (see section 6.5.2).

The justification for implementing the TAMA methodology on temporal features is the same as for

a/p features (section 7.2): previous evidence of temporal accommodation is primarily based on across-dialogue comparisons (Bosch *et al.* 2004b, 2004a; 2005), which do not reveal the occurrence of accommodation in a continuous representation. Exceptions to the above are studies on *rhythmic coordination* of speakers (Jaffe *et al.* 2001) and a recently published study (Edlund *et al.* 2009).

8.2.1 Annotation of switch pause and overlap

Floor exchanges between speakers are better visualized schematically by means of a chronograph, shown in Figure 8.1. In order to obtain a picture of the floor exchanges, the individual chronographs of the two speakers, which were obtained from the semi-automatic annotation process described in section 6.5.2, are added together. This results in a combined chronograph on which there can be only one of four situations: vocalization by speaker A, vocalization by speaker B, overlapping speech, and silence (pause).

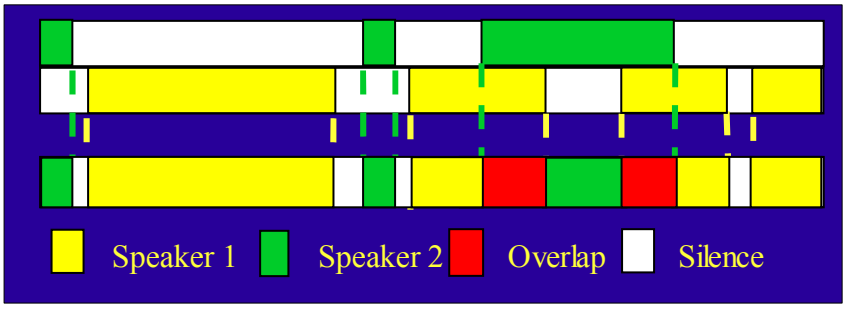


Figure 8.1: Part of dialogue chronograph for two speakers (individually and combined)

Numerous approaches to defining speaker turns from the combined chronograph have been proposed (Beattie 1982; Weilhammer and Rabold 2003; Bosch *et al.* 2004b; Benus 2009). As pointed out in (Raux and Eskenazi 2008) and (Bosch *et al.* 2005), this is difficult to do based on the chronograph itself. For example, the third utterance of speaker 2 (green) in Figure 8.1 could be characterized as a turn, or as a short non-interrupting contribution in an otherwise speaker 1 (yellow) dominated part of the dialogue.

In general, it is beneficial to be able to obtain a simple definition of switch pauses and overlaps from the chronograph for two reasons: (a) temporal information is *objective,* as opposed to discourse analysis based on *assumptions* about the speakers' intentions and meaning of utterances (Raux 2008), and (b) temporal information may be the only available data when analyzing large databases or performing online analysis (Bosch *et al.* 2004b).

Thus, one of the simplest possible schemes was proposed in (Kousidis and Dorran 2009) for characterizing switch pauses and overlaps and attributing them to either speaker (see Figure 8.2):

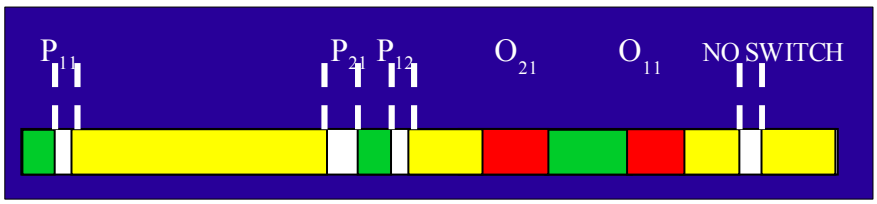


Figure 8.2: Switch pause and overlap definition and speaker attribution

Switch pauses occur between vocalizations that belong to different speakers. These are considered to "belong" to the speaker who takes the floor *after* the pause: **P_{11}** is a switch pause that belongs to speaker 1 (yellow) as the floor is given up by speaker 2 (green). The opposite occurs at **P_{21}**: yellow gives up the floor and, after a pause, green takes over. A similar rule is implemented for overlaps: the speaker who initiates a vocalization during an utterance of the other speaker is the "owner" of the overlapping segment: in **O_{21}** speaker 1 (yellow) is talking when speaker 2 (green) initiates an overlapping vocalization: this overlap is attributed to green, who keeps the floor *after* the overlap segment (the opposite occurs in **O_{21}**). If after the overlap segment the floor is not exchanged, i.e. the speaker who had the floor retains it, then the overlap is categorized as non-interrupting (no switch of floor) and belonging to the *other* speaker. The same is the case for a pause between two speech intervals of the same speaker, shown on the right-hand side of Figure 8.2, which is not a switch-pause.

The annotation of switch pauses and overlaps can be performed automatically by means of a simple algorithm (Figure 8.3): each interval in the combined chronograph has only three properties: its start time, end time and *label.* The label can either be "speaker 1", "speaker 2", "pause" and "overlap". Looping through all intervals, the algorithm identifies those labeled "pause" or "overlap". If such an interval is found, the two neighbouring intervals are compared. If these belong to different speakers, the pause or overlap is characterized as a switch pause or interrupting overlap and attributed to the owner of the second interval. If they belong to the same speaker, the non-switch pause is attributed to the owner of the two intervals and the non-interrupting overlap is attributed to the *other* speaker.

In the (extremely) rare cases where overlaps are adjacent to pauses, the following rules apply: overlaps followed by pauses are non-interrupting. *Simultaneous starts* (overlaps immediately after a pause) are non-interrupting overlaps. The speaker that keeps the floor after the overlap is considered as the initial floor holder and the non-interrupting overlap is attributed to the other speaker. The code implementation of this algorithm can be found in appendix C.

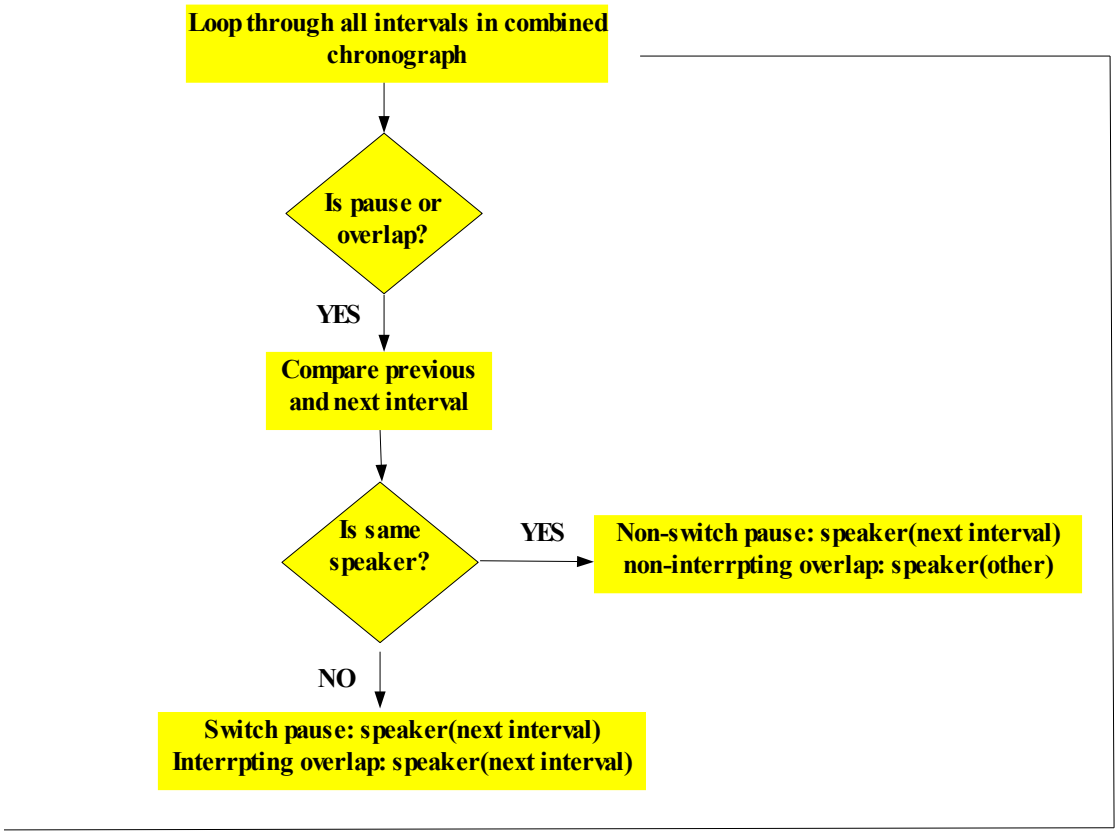


Figure 8.3: Algorithm for automatic annotation of pauses and overlaps based on combined chronograph

8.2.2 Feature average calculation

The annotation procedure described in the previous section yields four different measures for each speaker: The number of switch and non-switch pauses and the number of interrupting and non-interrupting overlaps. If the dialogue portion shown in Figure 8.2 is considered as a TAMA frame, then the *average pause length* (APL) for each speaker in the frame is given by:

$$APL = \frac{\sum_{i=1}^{N} d_i}{N}$$

Equation 8.1: Frame average pause length calculation

where d_i is the duration of pause i and N is the number of pauses attributed to that speaker in the frame. The same formula applies for switch pauses, non-switch pauses or both. Unlike the a/p features studied in chapter 7, the durations of pauses cannot be reasonably assumed to have a normal distribution around a mean value: there is a minimum pause threshold which is defined as a parameter in the silent/non-silent interval segmentation algorithm (6.5.1) and is typically 50-100 milliseconds (but shorter pauses can be introduced during the manual segmentation phase), but

there is no maximum. This results in a positively skewed distribution, in which large values bias the mean significantly. This can be overcome either by (a) taking the median value rather than mean, (b) setting a threshold above which all values are considered as outliers and are ignored, or (c) using a log transformation, e.g. $log_{10}d_i$,with d_i expressed in milliseconds (see Figure 8.4). The threshold in case (b) can be set by assuming an *exponential* distribution and removing all values with p<0.05 on the right-hand side tail.

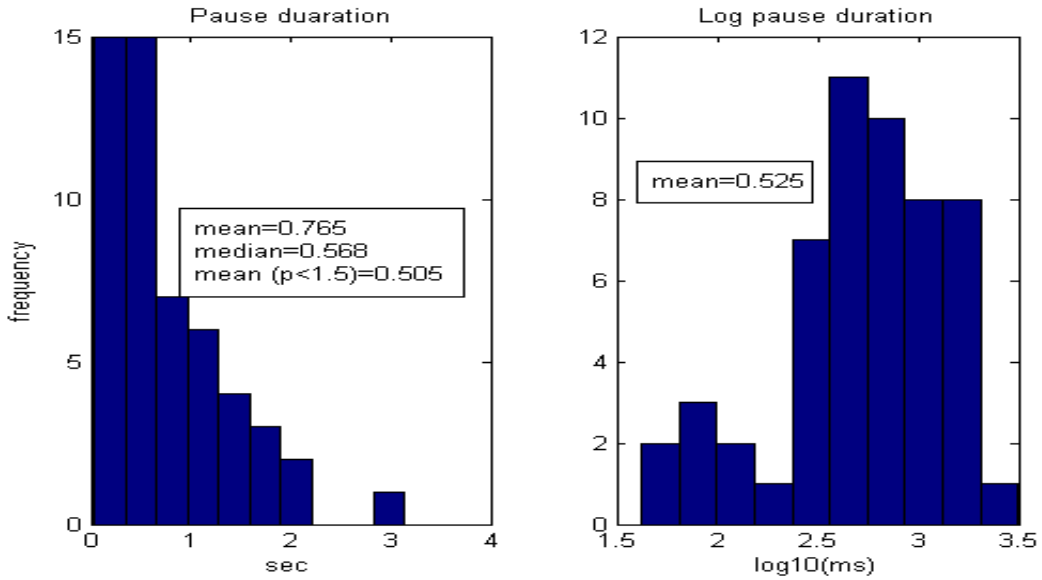


Figure 8.4: Histograms of pause duration distribution (left) and log duration distibution (right)

A different measure is defined for overlaps, which is termed overlap rate (OR). This expresses the amount of vocalizations initiated as overlaps over the total amount of vocalizations of that speaker, within a TAMA frame. In other words, how often a speaker tends to speak *before* the partner has finished their utterance:

$$OverlapRate = \frac{OverlapCount}{TotalCount}$$

Equation 8.2: Calculation of overlap rate

The overlap count may or may not include non-interrupting overlaps, in which case the total count is adjusted accordingly. In the former case, it is the total number of a speaker's vocalizations minus those occurring after a pause and OR expresses the tendency of a speaker to overlap in general. In the latter case, the total count is the number of vocalizations after a switch-pause or an interrupting overlap and OR expresses the tendency to take the floor by interrupting the other speaker.

It is noted that interruption does not necessarily imply a pragmatic function: in spontaneous speech it is quite common for speakers to take the floor before the interlocutor has finished their utterance

without this being considered an interruption. This is because interlocutors often understand each other without having to listen to the complete utterance, and are able to respond earlier thus increasing *efficiency*. However, this type of behaviour may not be considered polite in certain contexts (e.g. formal interviews), while in some cases it could be a sign of positive evaluation between interlocutors (friendly chat). Therefore accommodation/non-accommodation of OR is interesting from the point of view of SDS, as it could increase the "friendliness" of a talking agent.

Another important note involves non-speech intervals which were annotated separately in the corpus (see section 6.5.2). These include breath noises, instances of laughter and other non-speech sounds. Since these intervals *are* part of the temporal structure of the dialogue, they should be included in the analysis. For example, an audible breath before an utterance is a signal that a speaker is about to initiate a vocalization, and the partner is likely to interpret it as such. Similarly, laughter is a vocalization produced in response to a previous utterance and can be considered as interrupting overlap. On the other hand, laughter is "overlap-inviting", i.e. a speaker is more likely to overlap while the partner is laughing (in order to extend the joke) and this may bias the overlap rate, as instances of laughter are frequent in the corpus.

Finally, it is noted that, for these temporal features (APL and OR), the TAMA frame length trade-off is much more severe than it was for a/p features. Frame lengths of 30 seconds contain approximately 10 instances per speaker (sometimes less), and this number includes both pauses and overlaps. Therefore, the calculation of an average value is much less robust, and one large value may severely bias the analysis. The only solution is to increase the frame length which, as discussed in chapter 7 reduces the resolution of the TAMA representation. Therefore, there is a problem of data *sparsity* when using the TAMA method on temporal features.

8.2.3 Pilot study

A pilot study (Kousidis and Dorran 2009) was conducted in order to test the effectiveness of TAMA in describing inter-speaker accommodation of temporal features. The five dialogues from the "shipwrecked" corpus that were presented in (Kousidis *et al.* 2009a) were analyzed for temporal accommodation. The analysis focused on switch pauses and overlap rate including both interrupting and non-interrupting overlaps. Only speech intervals were considered in the analysis and non-speech intervals such as instances of laughter were excluded. The pause duration distribution was not transformed in any of the ways discussed in the previous section, but a small number (<3) of extremely long instances were excluded from the analysis of each dialogue, without setting a pre-defined threshold. The techniques described in the previous section were considered as a result of the pilot study. Application of all three techniques to the entire corpus (see appendix A) did not

contradict the findings of the pilot study.

An across-dialogue comparison showed a linear relationship between the APL of the two speakers (see Figure 8.5), confirming the findings of (Bosch *et al.* 2005). In the latter study, this linear relationship was attributed to two possible causes: (a) accommodation of APL between the two speakers, or (b) a result of the overall "liveliness" of the dialogue, as a more lively dialogue would exhibit silent intervals of shorter durations. This is discussed further in section 8.3.3.

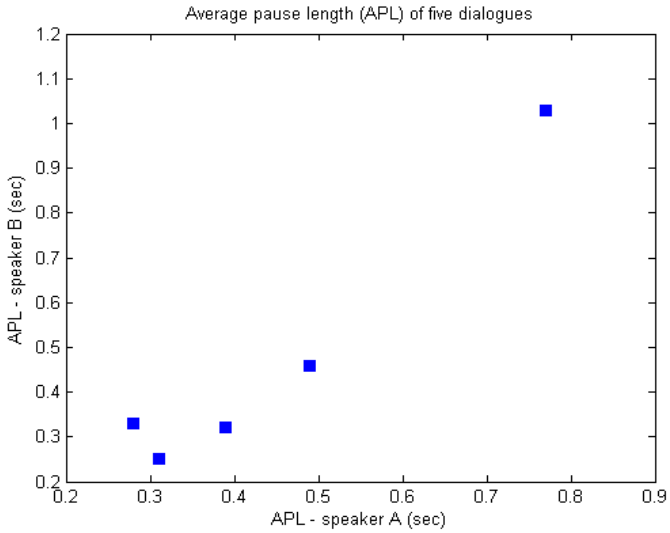


Figure 8.5: APL of speakers in 5 "shipwrecked" dialogues

In order to test the hypothesis of accommodation, Kousidis and Dorran (2009) implemented a TAMA analysis for both APL and OR (see Figure 8.6):

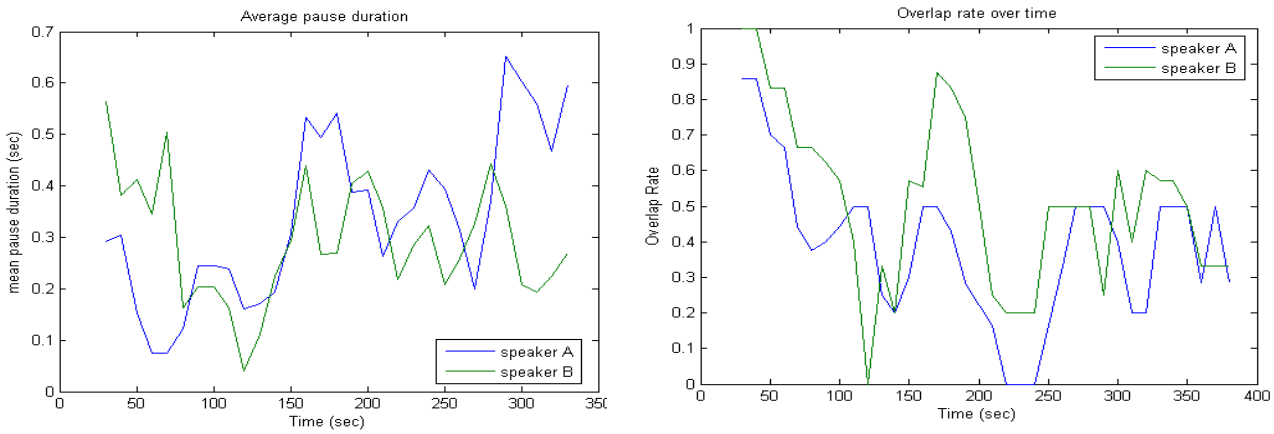


Figure 8.6: TAMA plots of APL (left) and OR (right) calculated over frames of length 30s with 33% overlap

In these plots (taken from different dialogues), a similar trend for both speakers is discernible; however, accommodation is not as evident for all five dialogues studied in (Kousidis and Dorran

2009) as shown in the examples of Figure 8.7. Even rarer are the cases in which the similarity can be evaluated statistically (see appendix A). Of course speakers' temporal features do not *have to* converge, as discussed in section 7.3.2; but the linear relationship of the overall dialogue average values that was also reported in (Bosch *et al.* 2005)suggests that accommodation is more ubiquitous than evidenced in the TAMA plots.

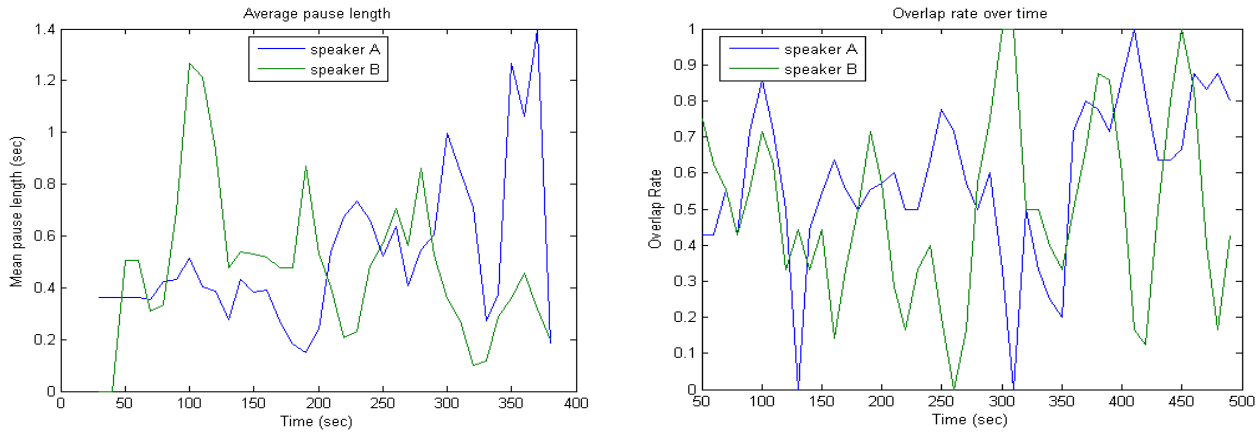


Figure 8.7: TAMA plots of APL (left) and OR (right) calculated over frames of length 30s with 33% overlap

The data sparsity problem that was mentioned in the previous section was identified as a possible cause for this lack of evidence of temporal inter-speaker accommodation. Thus, it is possible to increase the frame length in order to increase robustness, at the cost of resolution; but it is desirable to keep the resolution high, as this is one of the main advantages of TAMA. The upper limit for the frame length is the dialogue duration itself, which would degenerate the TAMA method into an across dialogue comparison method.

A later study (Edlund *et al.* 2009) which followed a methodology similar to TAMA, introduced frames of varying length by defining a window of 20 instances with an overlap of 19 instances (each window contained exactly 20 switch-pauses or 20 pauses). This resulted in frame lengths as long as 180 seconds. On the other hand, the resolution was kept high, as a new window was defined at each pause. This approach is similar to TAMA (increasing robustness by introducing overlapping frames), taken to the extreme (frame overlap equal to 95%). However, this would result in a time series with 19 significant coefficients in the correlogram, as each frame shares common instances with the previous 19 frames. This would make difficult to estimate model parameters using the methodology described in section 7.4.4. A possible approximation could be to assume exponentially decaying weights on the previous instances.

(Edlund *et al.* 2009) found results that were similar to those of (Kousidis and Dorran 2009): a

portion of the dialogues exhibited statistical evidence of contemporaneous accommodation in APL, while in other dialogues speakers' APL did not converge or even diverged. This does not necessarily imply that there is no temporal accommodation in the dialogues where the local averages do not follow similar trends, especially since (a) there is no *perceptual* difference when listening to the dialogues, and (b) the same dialogues that do not exhibit accommodation locally contribute positively to the linear trend found for across-dialogue comparisons shown in Figure 8.1.

An alternative explanation of the findings in (Kousidis and Dorran 2009), also given in (Edlund *et al.* 2009), was that the TAMA method is not sufficient in this case (of temporal accommodation) because the variation introduced by factors other than accommodation is relatively much larger than in the case of a/p features. Such factors could be specific dialogue modes, utterance types (e.g. back-channeling), or speaker dominance: if the dialogue is dominated (locally or globally) by one of the two speakers, this may have an effect on either (or both) speaker's APL and OR. Therefore, the value of APL and OR at any arbitrarily defined frame will be a function of various factors which introduce variations, so that accommodation is "masked", or, as proposed in (Edlund *et al.* 2009), "overridden". Another way to view this is that speakers accommodate their silence durations and overlapping speech behaviour but this process is not necessarily *synchronous:* the dominance factor, for example, suggests that speakers shorten their pause durations and increase their overlap rates (due to increased back-channeling) *when the dialogue is dominated by the partner* (a hypothesis), i.e. *not* contemporaneously. This is further discussed in section 8.3. Therefore, an important outcome of (Kousidis and Dorran 2009) is that an SDS strategy of accommodating to the user's APL synchronously (in order to optimize its end-pointing threshold) would be too simplistic, as it would disregard all other factors of variation in APL, therefore leading to unnatural behaviour of the talking agent.

In conclusion, although the TAMA method shows that accommodation occurs at a local level, it is insufficient in itself for the purpose of describing temporal accommodation, mainly because significant variation in temporal features is introduced by other factors (which are related to the discourse and require discourse analysis, e.g. floor exchanges), but also because of data *sparsity*, i.e. the small amount of feature instances in the TAMA frames. The inadequacy of serial approaches such as TAMA and the methodology of (Edlund *et al.* 2009) to capture temporal accommodation arises from the schemata of turn attribution that are based solely on the chronographic representation. Thus, a novel dialogue representation (turn-share distribution) was formulated in order to explore the effect of floor exchanges on temporal features. This representation is described in the next section.

8.3 Flexible dialogue representations

The schema for annotation of switch pauses and overlaps that was employed in (Kousidis and Dorran 2009) is not the only one possible. For example, a more complex schema is proposed in (Weilhammer and Rabold 2003), in which 10 distinct configurations of pauses and overlaps are used to describe the process of turn-taking in human dialogues. As pointed out in (Bosch *et al.* 2005), such rule sets are ambiguous, especially when attributing overlapping speech to one of the two speakers. For example, (Adda-Decker *et al.* 2008) defined four types of overlap in order to annotate a corpus of political interviews, but categorization of all instances to one of the four categories required a semantic and pragmatic analysis of the dialogues, which introduces a certain degree of subjectivity. The situation with most complexity is that of spontaneous dialogues, in which speakers barge-in "out of turn" without this being considered as an interruption: this can either be characterized as an interrupting short turn, if the floor is given back to the original speaker, or as a non-interrupting out-of-turn speech segment (i.e. not a turn).

Thus, the "half-duplex plus feedback" model (Figure 6.1) that was used in order to describe accommodation of a/p features is insufficient in describing temporal accommodation, as the temporal organization of dialogue does not comply to this schema: dialogue does not flow "back and forth" as a sequence of interchanging of turns: this is only a *representation* of the dialogue which has been dominant since the invention of the interaction chronograph[22] (Lennes and Anttila 2002), due to its intuitiveness and utility to a certain extent. However, other representations are possible, in which it is not necessary to define "turns". Such a representation is presented in the next section.

8.3.1 Turn share

In order to overcome the issues discussed in the previous sections, (Kousidis *et al.* 2009b) proposed a new dialogue representation which completely ignores the notion of "turns" and replaces it with a new measure termed *turn share*. This section describes the dialogue representation proposed in (Kousidis *et al.* 2009b).

Let the part of the dialogue shown in Figure 8.1 be a frame of known length, **L**. During this time frame, both speakers share the floor at any time, both when they are speaking as well as when they are silent. The dialogue is a shared experience where each interactant participates with their speech, but also their choice to remain silent and actively listen, instead. This assumption is consistent with

22 The interaction chronograph was a mechanical device (basically a typewriter with continuous paper feed) which could record the times of events in a dialogue by means of key strokes (Lennes and Anttila 2002; Campbell 2009)

the viewpoint of synchronous interaction (Campbell 2009). Therefore, at any time point in the frame, each speaker can be in only one of two states: active (speaking) or passive (silent). "Speaking" may also include non-speech segments, as the speaker is active when producing them. These "states" are expressed by two proportional measures, *active time AT* and *passive time PT* as:

$$AT = \frac{L_A}{L}, PT = \frac{L_P}{L}$$

Equation 8.3: Definition of active (AT) and passive time (PT) as proportions of vocalization and silence in a frame of length L

where L_A, L_P are the total durations of speech and silence in the frame, respectively. These two measures have the property $AT + PT = 1$, which means that one can obtain AT by annotating only the silences and calculating $AT = 1 - PT$. The turn share TS is then defined as:

$$TS_1 = \frac{AT_1}{AT_1 + AT_2}, TS_2 = \frac{AT_2}{AT_1 + AT_2}$$

Equation 8.4: Definition of turn share

where TS_n, AT_n are the turn share and active time of speaker n, respectively. Apart from the obvious property $TS_1 + TS_2 = 1$, the definition can also be extended for interactions with more than two speakers. Also it should be noted that $(AT_1 + AT_2)$ can be (and often is) longer than L, the length of the frame, due to the overlaps. In order to comprehend the physical meaning of turn share, it is helpful to look at a plot of turn shares over time (see Figure 8.8)

Figure 8.8: Turn share plot obtained by calculating TS for two speakers over 4-second-long frames

In this plot, each bar corresponds to a frame four seconds long and the vertical axis (time) progresses from the bottom to the top of the figure. Only a part of the dialogue is shown (approximately 4.5 minutes long). In this representation, it can be seen whether one of the speakers dominates the dialogue and in which parts. This can be useful for statistical analysis of temporal and other features depending on whether a speaker has a greater turn share, as opposed to whether it is his/her turn. The plot shows turn shares calculated for adjacent non-interrupting frames, but the representation can also be implemented as a continuous, *real-time* indicator, as shown in Figure 8.9: the colored bar boundary moves along the horizontal axis continuously as turn share is calculated for the previous *n* seconds, where *n* is a fixed value.

Figure 8.9: A continuous indicator of turn share

The optimum frame length of the representation depends on the application. Shorter frames reveal a finer picture of the interaction in terms of turn-share exchange. In fact, taking progressively shorter frames leads to more instances of frames that are completely dominated by one of the speakers, indicated by uniform color bars that extend from one end the other in the turn share plot (bars extend from left to right for one speaker and from right to left for the other speaker). The *limit* of the representation (i.e. infinitely short frames) is the chronograph of Figure 8.1 and, in particular, the individual chronographs for each speaker. Longer frame lengths lead to less instances of totally dominated frames and are likely to include complete utterances, which is useful for analysis of a/p features. The disadvantage of longer frames, similarly to TAMA frames, is loss of resolution and the danger that an equally shared frame may in fact be two adjacent half-frames that are totally dominated by each speaker. This issue is further discussed in section 8.3.2.

In addition, it would be desirable to obtain a representation for the amount of overlap and silence in a given frame. Considering Figure 8.2, four measures (proportional to the frame length, L) can be defined in similar manner to the definition of AT and PT above. These are shown in Table 8.1 below.

Measure	Descriptiom	Formula
S1	Speaker 1 portion	$S1 = L_{S1} / L$
S2	Speaker 2 portion	$S2 = L_{S2} / L$
TP	Total silence	$TP = L_{P} / L$
TO	Total overlap	$TO = L_{O} / L$

Table 8.1: Definitions of proportions in frame for speaker share, overlap and silence

where L_{SN} is the total duration of speaker N, L_O is the total duration of overlapping speech, and L_P is the total duration of silence in the frame.

It follows from the definition that the sum of all four proportions equals one, and that the quantity *(1-TP)*, hereafter *joint active time, JAT,* is the sum of the other three proportions. JAT is a measure of how "engaged" or "active" (in terms of liveliness or for example in presence of a debate) the particular part of the dialogue is, as more active dialogues are expected to have shorter pauses. A similar measure of liveliness, used in (Jaffe *et al.* 2001) is the vocalization over pause ratio *V/P* which is positively correlated to $AT = V/(P+V)$ (for individual speakers), and JAT (both speakers simultaneously). The overlap time, *TO,* is also expected to be a good indicator of high activation, as it is expected to be positively correlated to JAT. Therefore, a direct application for these two quantities may well be automatic recognition of activation (in the context of emotional speech) in spontaneous dialogues.

Figure 8.10 shows a per frame turn distribution plot for one of the dialogues recorded. As in Figure 8.9, the vertical axis (time) progresses from the bottom to the top of the figure. The red areas are stretches of the dialogue characterized by large amounts of overlap speech (non-speech intervals are included). JAT equals the length of the bars (from left to right), as the silence proportion TP is drawn in white. One can discern that longer bars seem to coincide with red areas (overlaps), which implies that more active speech is characterized by longer and/or more frequent overlaps. If a frame length equal to the duration of the entire dialogue is used, then the representation yields a turn share distribution for the entire dialogue, previously presented in (Lennes and Anttila 2002)

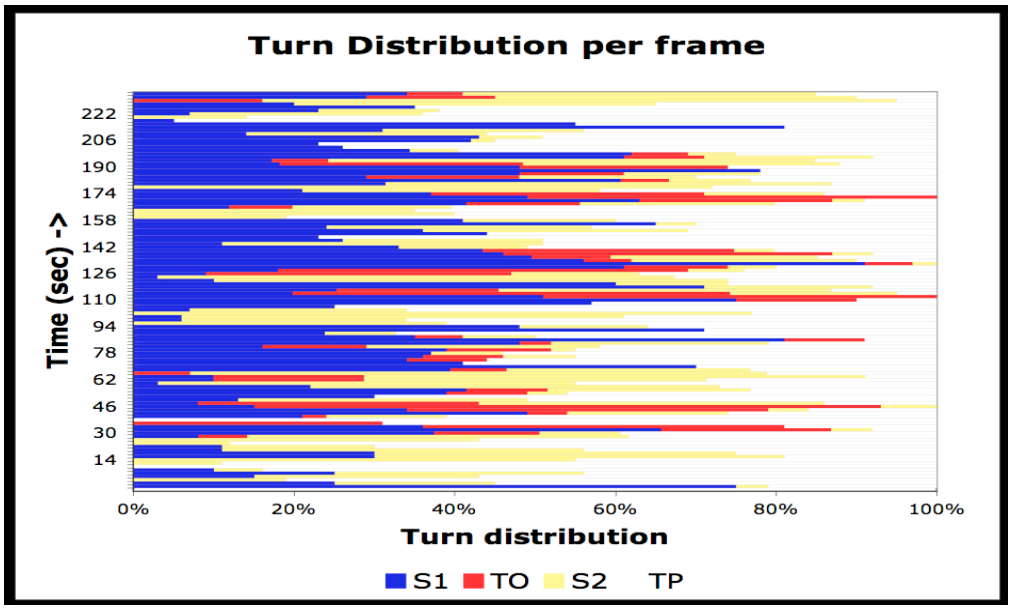


Figure 8.10: Per frame turn distribution for frame length equal to 4 seconds (50% overlap)

8.3.2 A practical example

In order to test the usefulness of the proposed representation, (Kousidis *et al.* 2009b) carried out a preliminary analysis, based on 5 dialogues recorded using the "shipwrecked" scenario experimental setup. In this study, the speakers' average pause length (APL) and overlap rate (OR) were investigated in relation to turn share (TS) and joint active time (JAT) distributions. Since the proposed representation does not define turns for the speakers, the pauses and overlaps were attributed to the speakers in an unambiguous manner (see Figure 8.2): pauses belong to the speaker who initiates a vocalization immediately *after* the pause interval, regardless of who is speaking before the pause; for overlaps, the interval immediately before the overlap segment, is considered, and the overlap is attributed to the speaker who is *not* speaking in that segment (thus initiating the overlap segment). There is no distinction between switch and non-switch pauses, or interrupting and non-interrupting overlaps.

The results of the study in (Kousidis *et al.* 2009b) are shown in Table 8.2. A frame length of 5 seconds with no overlap was used. APL was found to be strongly correlated to JAT. The correlation is negative, which indicates that high JAT results in shorter pauses. This is intuitive to an extent, as JAT is defined as the proportion of vocalization (total length minus the total duration of silence). Thus, it is possible that there are fewer – and longer – pauses. This correlation validates the hypothesis that there are in fact *shorter* pauses. OR is positively correlated to JAT, which indicates that high JAT results in more frequent overlaps. Again, OR is positively correlated to TO, but expresses the *frequency* of overlapping segments, rather than their relative length (TO). This finding validates that overlaps are more *frequent* when JAT is high. OR is also (negatively) correlated with TS, which indicates that speakers overlap their interlocutors more often when they have a smaller turn share (e.g. due to back-channeling).

Dialogue		TDD (sec)	APL JAT	APL ER	OR JAT	OR TS
1	F	428	-0.6	-	0.3	-
	M		-0.5	-0.3	0.5	-0.3
2	M	490	-0.6	-0.3	0.4	-0.4
	M		-0.7	-	0.5	-0.2
3	M	409	-0.6	-0.3	0.5	-
	F		-0.4	-	0.6	-
4	F	516	-0.5	-0.5	0.4	-0.4
	F		-0.7	-0.4	0.4	-0.3
5	M	363	-0.7	-	0.4	-0.4
	M		-0.4	-0.3	-	-

Table 8.2: Correlation coefficients between APL, OR and JAT, TS and ER (Significant at 95%, t-test with n-2 degrees of freedom, where n equals the length of the data).TDD: Total Dialogue Duration

A correlation between APL and turn share, TS was not found. However, APL is correlated with a related measure, hereafter *exchange rate,* ER, defined as ER = $2 \cdot MIN(TS_1, TS_2)$. ER takes values between 0 and 1 and expresses the degree to which a frame is dominated by either speaker (zero) or shared (1). The correlation between APL and ER is negative, which suggests that speakers shorten their pause length when exchange rate is high, i.e. when the floor is shared more equally.

The results of (Kousidis *et al.* 2009b), support the argument that the proposed representation of spontaneous dialogues can be useful in verifying the effect of factors such as JAT and ER on temporal features. One advantage of this representation is that it moves away from turn attribution and, consequently, the shortcomings of defining turns solely from the chronograph of the dialogue. Clearly, meaningful turn segmentation can only be achieved by discourse analysis which, in the context of SDS, pre-requires automatic speech recognition (ASR) and spoken language understanding (SLU) output. However, it is desirable for the interaction management component (which manages when the system can speak to the user or when the user's turn has ended) to operate independently of these components, due to their higher computational load and significant error rates in practice. For this reason, spoken dialogue systems have to rely on low-level information from the signal to manage turn-taking behaviour, namely the duration of turn-switch pauses and prosodic features such as final vowel lengthening. The approach presented here provides an alternative solution: the interaction management component can adapt to the *ongoing* session and adjust its thresholds and latencies according to JAT and ER. It would be naive to consider that the methodology outlined here could *replace* the current methods of SDS design; rather, the proposed representation should at best be seen as a starting point towards more flexible representations of the dynamics of human (and human-computer) interaction, which in turn may push naturalness of SDS forward.

One argument against the representation presented here is that there is loss of information due to the averaging "sliding window" process. Indeed, the length of the applied frame determines the time resolution of the representation. But, as indicated by the example analysis presented in (Kousidis *et al.* 2009b), there is nothing preventing the use of the original chronograph in order to extract features and analyze them *in combination* with the proposed representation. The purpose of the averaging is only to extract information about the turn-share distribution properties *in the neighborhood* of a segment (in this case a pause or an overlap). This information can also be combined with other inputs, such as low-level acoustic and prosodic features.

The size of that neighborhood, or frame-length, is another feature that needs to be considered. As discussed earlier, there is a trade-off between time resolution and frame length. It is desirable to keep the frames (and consequently the time resolution) small, because ER (or TS) is very sensitive

to frame length: the worst-case scenario is that a frame with ER=0.5 is actually two adjacent "half-frames" with ER = 1 (each speaker dominating one of the adjacent half-frames, yielding an equally shared frame). This can be allowed for short frames, because even when this is the case, responses are often anticipated *before* they occur, therefore the speakers *know* that there is going to be an exchange. Indeed, the correlations in Table 8.2 remain significant for frames with length approximately up to 8 seconds. JAT and TO, on the other hand, are less sensitive to frame length, and can be used to monitor lower frequency variations in activation, or engagement in the dialogue.

Another important point is that APL is correlated to JAT and ER, which apply to both speakers equally at any time in the dialogue, although each speaker's APL is influenced differently by JAT. Therefore, the proposed representation did *not* reveal a source of variation in APL that would imply non-contemporaneous inter-speaker accommodation (see section 8.2.3). This was the case however for OR, as it was found to be correlated to TS, therefore a lag zero correlation of the two speakers' OR should not be expected unless the dialogue (or part of) is characterized by high ER, in which case turn shares tend to be equal most of the time.

Finally, considering turn shares rather than turns is more consistent with dialogue representations which consider both speakers active at any time during the dialogue (Campbell 2009; Heylen 2009). Thus, the dialogue schema of Figure 6.1 can be updated in order to represent this view. In a full-duplex model, properties of speech are not necessarily causally related to the immediately preceding time interval in the interaction, but subject to the ongoing interaction in which both speakers participate equally. The process of instantaneous feedback that was discussed in section 7.5 is one aspect of this: a/p and temporal (and possibly other) features of speech are subject to variations at the instant the feedback is perceived, i.e. *during* vocalization and not *after*. The simplest possible way to depict this process is to superimpose Figure 8.9 on Figure 6.1 resulting in the following schema, which is equivalent to the schema proposed in (Heylen 2009)

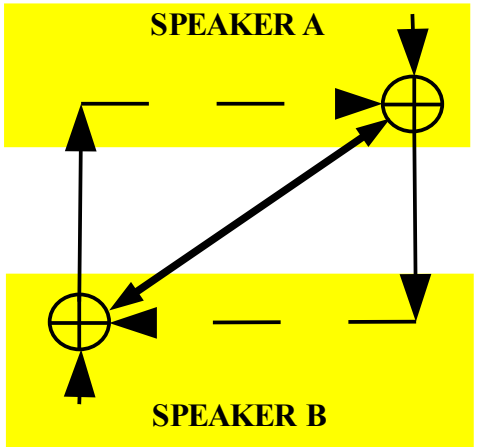


Figure 8.11: Representation schema of dialogue including instantaneous feedback

8.3.3 Accommodation or liveliness: a case study

This section describes further work carried out in order to answer the question introduced in section 8.2.3: is the correlation between the average pause lengths of two speakers the result of inter-speaker accommodation, or a result of the overall dialogue liveliness? The findings in (Kousidis *et al.* 2009b) suggest the second hypothesis, as APL is correlated to JAT. However, this does not imply that JAT is the *only* source of variation in APL.

This study was based on a corpus of 34 telephone dialogues in Japanese (Campbell 2009). The average duration of these dialogues is approximately 30 minutes, and the annotation comprises a chronographic segmentation for each speaker, in which speech, silence, and non-speech intervals are separately labeled. The dialogues were split into four quarters (duration equal to approximately 7.5 minutes) in order to deal with the data sparsity problem discussed in section 8.2.2. In addition the *threshold method* was used to deal with the skewness of the pause length distribution (see section 8.2.2). A threshold of 1 second was found to be reasonable based on the actual distribution (less than 2% of all pauses were above this threshold).

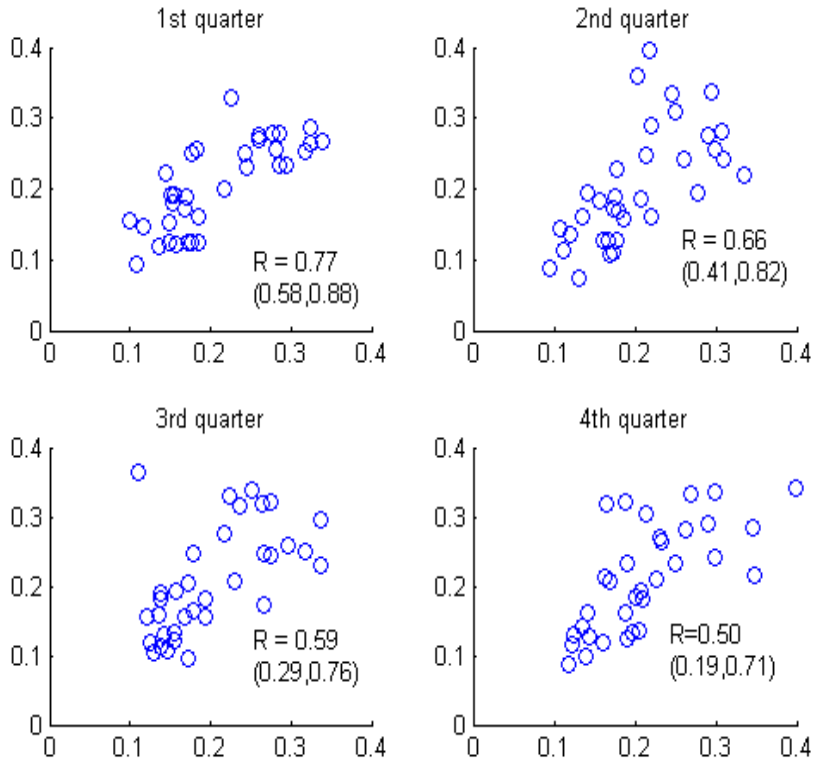


Figure 8.12: Correlations between APL of the speakers in 136 dialogue parts equally binned per order in time. X & Y axis in seconds. *$p<0.01$ for all coefficients (t-test with n-2 degrees of freedom where n equals the length of the data, confidence intervals in parentheses)*

The quarter split resulted in 136 dialogue parts, which were in turn divided into 4 equal sized bins of 34 parts, based on two conditions. First, according to their position in the dialogue (i.e. initial, second, third, final), and second according to the JAT. For each quarter, the zero-lag correlation coefficient for the two speakers' APL was calculated.

Figure 8.12 shows the results for the condition of order in time. The average pause lengths of the two speakers are positively correlated in all 4 quarters. This constitutes a statistical evaluation that speakers accommodate to each other's APL over time, although the resolution is very low due to the data sparsity problem. The p-values are low (<0.001) for all quarters, which implies that the frames are too long and that the optimal frame length for providing evidence of continuous temporal accommodation is less than 7 minutes (i.e. higher resolution can be achieved). This would however result in less points for each frame, further widening the confidence intervals. Interestingly, there is an apparent progressive declination in the strength of the correlation, although the confidence intervals are not narrow enough to validate this.

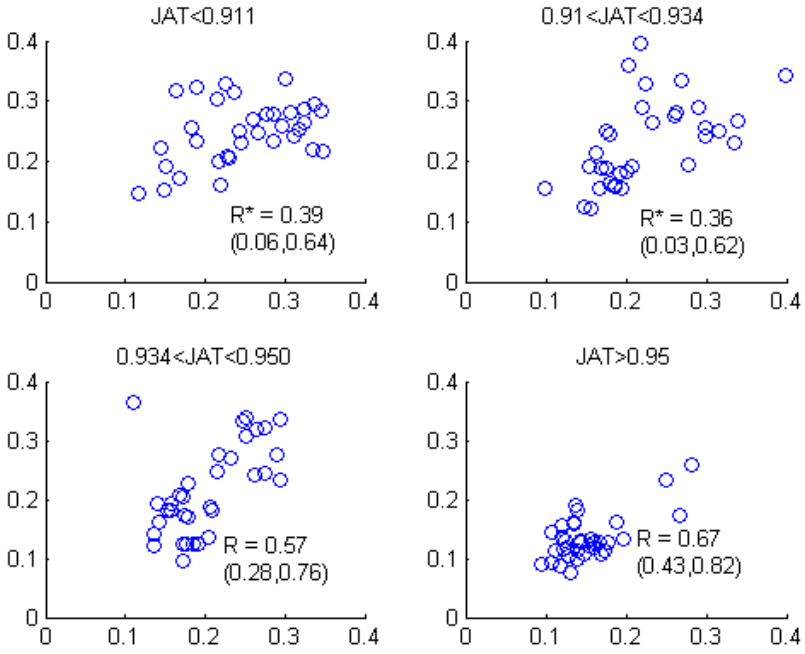


*Figure 8.13: Correlations between APL of the speakers in 136 dialogue parts equally binned per JAT. X & Y axis in seconds. *p<0.05, p<0.01 for all other coefficients (t-test with n-2 degrees of freedom where n equals the length of the data, confidence intervals in parentheses)*

In the JAT condition (see Figure 8.13), it was found that correlations between the two speakers' APL remain significant *regardless* of JAT. However, the results show that JAT (a measure of dialogue liveliness) is *not* the only source of the correlation: first, the APL does tend to become smaller as JAT increases, but the values are distributed over approximately the same range in all cases. In other

words, dialogues with *lower* JAT may have significantly *lower* APL, despite the fact that JAT and APL have a strong negative correlation (see previous section). Again, there is an apparent progressive *increase* in the strength (and significance) of the correlation as JAT increases, although the confidence intervals are not narrow enough to validate this.

These two findings provide sufficient evidence that the correlation between the speakers' APL can be safely attributed to inter-speaker accommodation. The alternative explanation of dialogue and topic liveliness (Bosch *et al.* 2005; Benus 2009) is not contradictory: the dialogue liveliness comes as a *result* of the interaction between the two speakers, whose speech is the only *input* to the process. Liveliness is not a third, external causal factor but an inherent property of the dialogue. An identical explanation was given in the case of a/p features (see section 7.5). This holds true even if liveliness is stimulated, e.g. by means of MIP experiments. Speakers may respond to such a stimulus, but the degree to which they respond and the effect that this response has on the APL for example, is determined by the interaction between them. Otherwise, it would be expected to find instances were the APL of the two speakers is not close to the "line" in the plot, but such points are extremely rare (one point in lower left plot in Figure 8.13). The linear relationship of APL found in various studies (Bosch *et al.* 2005; Kousidis *et al.* 2009b) implies that accommodation of APL is ubiquitous, but, due to the complexity of the temporal organization of dialogue, difficult to capture with continuous methods such as TAMA or the similar method in (Edlund *et al.* 2009).

8.4 Discussion

This chapter has presented two distinct approaches to describing inter-speaker accommodation of temporal features, namely a modification of the TAMA methodology described in chapter 7 and a novel dialogue representation based on turn shares and joint share distributions. The two representations are complementary: the TAMA methodology describes contemporaneous accommodation across speakers, while the turn share representation describes accommodation of each speaker towards dialogue activity as expressed by turn share (TS), exchange rate (ER) and joint active time (JAT).

Contemporaneous adaptation is not evident in the TAMA plots (Figure 8.7) and cannot be validated statistically in most cases. However, a portion of the dialogues in (Kousidis and Dorran 2009) shows remarkably similar variation in average pause length and overlap rate (e.g. Figure 8.6). Similar results were reported in (Edlund *et al.* 2009), where some of the dialogues showed synchronous variation in pause and gap length across the two speakers. This does not necessarily imply that no accommodation occurs in the other dialogues; rather, the inadequacy of these serial

approaches to capture temporal accommodation consistently arises from other sources of variation that are superimposed on the synchronous pattern (such as whether the dialogue is dominated by either speaker), as well as several other factors that serial analyses do not take into account, such as the different dialogue act categories.

A significant factor that influences TAMA analysis or other similar approaches is that of data sparsity: unless frames are long enough to contain tens of instances, the averages are biased by large values and the image is "blurred". But frame lengths of 3 minutes or more are closer to across-dialogue comparisons than to a continuous representation. The series are over-smoothed, and the apparent similarity in pause length is not more meaningful than an across-dialogue comparison, especially in view of the fact that some dialogues last less than 3 minutes. In (Jaffe *et al.* 2001), an optimal lag of 25-30 seconds was reported for rhythmic coordination between mothers and infants. However, this coordination was measured on the durations of five vocal states (see section 4.5.6) and again only a portion of the interactions showed coordination in pause length. Frame lengths of 30 seconds contain too few instances of pauses to calculate a robust average value.

Another possible cause of the inadequacy of serial methods is the annotation of switch-pauses and turn attributions solely from the chronograph of the dialogue. The categorization of pauses and switch-pauses – or pauses and gaps in (Edlund *et al.* 2009) – is probably inadequate in describing the temporal accommodation of natural dialogue. Such a representation might suffice for half-duplex interactions of the sort found in SDS environments, where there are clearly defined turns (as in request-response utterance pairs). In such scenarios, these categorizations can be useful. For example, an SDS may adapt its end-pointing threshold and latency according to the user's response latency. The time frame for this type of adaptation may be sufficiently long (1-3 minutes) in order to ensure robust behaviour, resulting in a slowly adapting system. However, this strategy would be inadequate for a system that could engage in free-from conversation.

An alternative approach was presented in (Raux 2008), in which the system based its turn-taking strategy on incremental analysis of prosodic, semantic and discourse structure information. The system made a decision as to whether a silence from the user was indicative of the end of their turn or not. Although this was implemented in a half-duplex interaction task (bus timetable system), some aspects of the analysis can be useful in order to describe temporal accommodation in spontaneous dialogues. In particular, a more informed categorization of pauses based on prosodic, semantic and discourse structure factors is likely to yield more informative results on synchronous accommodation, similarly to introducing dialog act classification for the a/p features: an overall average for each pause category is calculated for the whole dialogue, and each individual pause is

given a z-score compared to its category average, prior to calculating a mean z-score for the frame. Similarly, the occurrence of overlaps can be calculated as an average of z-scored probabilities of an overlap occurring at the actual overlap occurrences. This is perhaps the most promising path for serial approaches to describing temporal inter-speaker accommodation.

The turn share representation addresses the issue of turn attribution from a different point of view, totally disregarding turns and replacing them with turn shares. This is not a magical invention, but a simple mathematical formulation of a different perspective. Vocalizations of the two speakers are considered as occurring simultaneously, rather than in succession, in accordance to proposed *synchronous* descriptions of human interaction (Campbell 2009; Heylen 2009). The proposed approach can be used to model temporal behaviour on the turn share distribution of the current frame, as indicated by the strong correlations shown in Table 8.2. Thus an SDS could adapt its threshold based on the level of activity in the dialogue, for example by monitoring JAT. This could complement the approach in (Raux 2008) and further reduce the latency of the dialogue system responses.

The correlation between APL and OR with TS (and ER) is sensitive to the frame length, as the latter two measures are only meaningful if the dialogue frame is reasonably short (shorter then 8-10 sec). Otherwise, it is possible that a "shared" frame (high ER, or equal TS) is in fact a concatenation of two frames dominated by either speaker, in the worst case scenario. Thus, there is an optimal frame length, similar to the "optimal lag" of (Jaffe *et al.* 2001), in which the variation due to either speaker dominating the dialogue is most significant. In contrast, the TAMA approach, as well as other approaches (Bosch *et al.* 2005; Edlund *et al.* 2009) are more robust when considering longer frames or even entire dialogues, due to the data sparsity problem. Therefore, these can be seen as "macroscopic" approaches, while the proposed representation is meaningful only when applied *locally* (short frame length), and can thus be seen as a "microscopic" approach. A combination of the two is another possible route for extension of the work described here. The same holds true for other microscopic approaches, such as studies on rhythmic entrainment (Jaffe *et al.* 2001; Benus 2009).

In addition, the proposed representation provided evidence that the correlation of pause length between speakers across dialogues is not (solely) the result of higher/lower "liveliness" in the dialogue. Firstly, as was also discussed in section 7.5, the overall liveliness of the dialogue is not an external influencing factor but a *result* of the interaction, thus making the point moot. In addition, the analysis of the 34 telephone dialogues showed that values of APL are spread over a range of 100-400 ms for the same JAT, which indicates that the correlation across dialogues is indeed evidence of accommodation (since there are no other external factors). These findings support the

argument that the proposed turn-share representation is useful and constitutes a first-step towards other flexible dialogue representations which may provide useful insights in describing temporal accommodation.

9 Implementation of accommodation in SDS

9.1 Overview

This section describes the implementation of a Wizard-of-OZ SDS environment with accommodating behaviour in order to evaluate (a) whether this stimulates accommodation from the user, and (b) whether the user perceives the interaction as more natural in comparison to a control condition in which the system is not accommodating. This implementation comprised accommodation of a/p features, for which a sufficient model of accommodating behaviour was estimated in section 7.4.4. As discussed in section 5.3, the motivation for implementing accommodation in SDS arises from the need for more natural interaction between user and system. The procedure described here is consistent with the human metaphor paradigm (Edlund *et al.* 2008), which was presented in section 2.2.4.

It is noted that the work described this chapter constitutes a preliminary indicative approach: a fully operational implementation of inter-speaker accommodation in SDS (Wizard-of-OZ or actual system) lies outside the scope of this research, as the time commitment and resources required would classify such an endeavour as a separate project.

9.2 Design considerations

Two main design principles were considered in order to set-up a test platform for incorporating accommodation in SDS. The first one was that the SDS environment should be able to engage in "free-form" conversation with the user. This requirement arises from the fact that the description of accommodation and its statistical evaluation that were derived in chapter 7 were based on unconstrained, spontaneous dialogues. The characteristics of that description, namely the TAMA methodology, are more suitable to describing this type of speech than more constrained forms, for which better descriptions exist. For example, answer-question pairs or "form filling" tasks could be dealt with by utterance-based descriptions such as (Nishimura *et al.* 2008). This does not imply that TAMA cannot describe accommodation phenomena in such dialogues; however, a constrained dialogue task would not suffice as concrete proof that users accommodate their speech features to those of a system in more general cases: unconstrained dialogue is the most general case of speech, thus an implementation and evaluation of accommodation in an unconstrained human-machine interaction scenario is arguably more "powerful" evidence of the usability of the findings of chapter 7.

The second design principle was that the interaction task for this experiment required that the user and the system should have an equal role. The underlying motivation for this design choice is again

the experimental foundation of the human dialogues on which the description of accommodation is based and the fact that talker role has been found to influence accommodating behaviour (Brennan 1996; Pardo 2006). This adherence to equal conditions in the two settings (human-human and human-computer) is dictated by the principles of the human metaphor evaluation paradigm (Edlund *et al.* 2008), as comparisons between the two conditions are more meaningful when all other variables are kept constant.

In the human-human condition (chapter 7), accommodation of a/p features was found for dialogues in which two participants were cooperating in order to solve a task, namely the "shipwrecked" scenario(section 6.4.3). Therefore, it was decided to use this scenario in the test platform, in order to support the equivalence of the two conditions. All conditions of the experiment were the same: the human user and the computer agent had to cooperate in order to rank the 15 objects shown on screen in order of importance within 10 minutes (after this time the screen automatically turned off). Each subject would participate in at least two randomly ordered sessions, one of which would comprise a non-accommodating computer voice, while the latter would comprise accommodation along four dimensions: pitch, intensity, pitch range and speech rate. This could be further expanded into several sub-conditions, comprising accommodation of the system either along one dimension, all dimensions or any other combination thereof. A different scenario variation would be used for each session/subject from the three available: "shipwrecked", "space-pod" and "Himalayas" (see section 6.4.3 and appendix A). Subjects who had participated in the human dialogue recordings (mostly Digital Media Center staff) were excluded from this experiment, due to their knowledge of the task and the purpose of the experiment in general.

9.3 Technical implementation

The need for unconstrained dialogue between user and system dictated a Wizard-of-Oz implementation, in which human users situated in a soundproof isolation booth (see section 6.3) interacted with a hypothetical SDS, while they were explicitly told that they were talking to a fully automated intelligent system. The latter was implemented as a type interface, which the experimenter used to provide input to the TTS voice. Since the subjects could only hear the voice from inside the booth through headphones, there was no indication that could compromise their belief, other than the apparent "intelligence" of the system. The system voice introduced itself as "Kevin" and explained the task to the subjects, as they were initially unfamiliar with the task

The TTS voice used was FreeTTS[23] (version 1.2), an open source diphone voice synthesizer based

23 http://freetts.sourceforge.net/docs/index.php

on the Festival[24] speech synthesis system and a modified version of the FreeTTS Player[25] demo (see Figure 9.1). This interface comprises a text-box (bottom of panel) in which the wizard (experimenter) could type an utterance and instruct the application to play it by pressing the "Speak Text" button. In addition, typed utterances are stored in the play list (middle of panel), where they can be selected and played using the "Play" button.

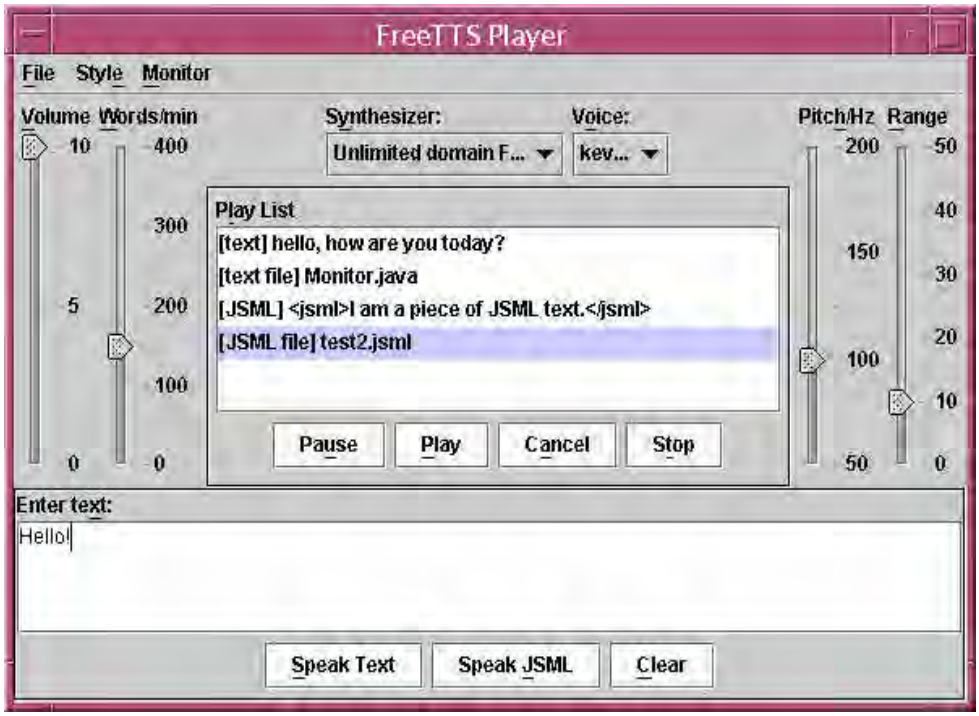


Figure 9.1: The FreeTTS Player interface

The four sliders (two on each side of the panel window) could be used to adjust the a/p features of the synthesized voice (pitch, intensity, pitch range and speech rate). The source code of the application was modified so that these sliders would be automatically adjusted whenever either the "Speak Text" or "Play" button were pressed, according to the values found in a text file. This file was updated every 10 seconds by a simultaneously running Praat script, which performed on-line prosodic analysis of the user utterances. The overall operation is shown in Figure 9.2. The user channel was recorded on a workstation by a real-time recording application in 10 second increments, thus producing an audio file (WAV format) every 10 seconds. The audio file was then loaded by the Praat script which performed the segmentation and feature extraction process described in sections 6.5.1 and 6.5.3. The script thus monitored the user average values per 10-second frame for each of the four a/p features, for which it calculated the normalized value (divided over the overall mean). It then calculated an updated normalized value for the system based on a

24 http://www.cstr.ed.ac.uk/projects/festival/
25 http://freetts.sourceforge.net/demo/JSAPI/Player/README.html

simple VAR(1) model of the form shown in Equation 7.8. The lag terms where preset as 0.7 for autocorrelation and 0.3 as a feedback term (a moderate value). This process was performed separately for each feature.

The resulting updated values for the a/p features were saved in the update file, which was then read by the TTS player upon request of "speaking" an utterance. The overall delay of the analysis was approximately two seconds. Therefore, the system voice accommodated its a/p features based on the previous 10 seconds of dialogue with a 2 second delay: Unless the system was required to speak within that period, then its features were "up-to-date", according to the simple VAR(1) model. However, the *current* interaction frame was not taken into account, and utterances generated towards the end of the 10 second frame carried a/p features that were adapted to the previous frame, thus missing up to 12 seconds of immediately preceding context in the worst case (10 seconds frame length plus 2 seconds for the delay).

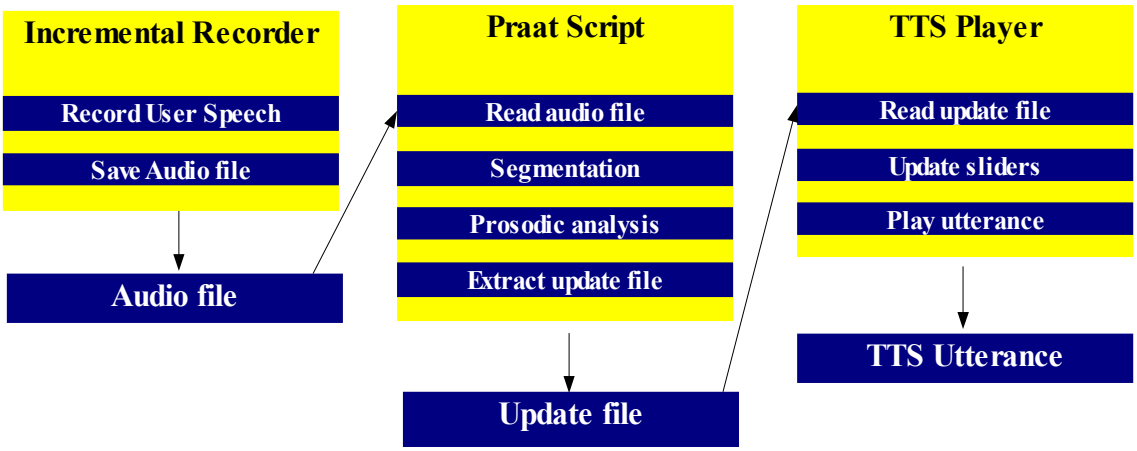


Figure 9.2: Schematic of operation for online analysis and TTS voice adaptation

Both the speaker and the TTS voice were simultaneously recorded on the ProTools console as described in section 6.3, providing high audio quality recordings for further analysis.

9.4 Performance

Upon initial testing of the testing platform, several performance issues were identified. First, the segmentation process (silence vs non-silence), as described in section 6.5.1, is *semi-automatic* and requires manual corrections, which were not possible in an on-line system. In addition, *annotation* of the intervals into speech and non-speech was not possible either. Therefore, a certain amount of error was introduced by mis-segmentation and mis-annotation of intervals. This error may introduce inaccuracy in the prosodic analysis and/or extreme values which would lead to erroneous accommodation of the system voice.

In order to overcome these problems, an optimized version of the Praat segmentation algorithm was implemented. The difference between the two algorithms is that the Praat function uses an intensity threshold based on the *maximum* intensity of the sound, while the optimized version uses a *preset* value as a threshold. With the Praat version, the optimum threshold is different for each frame, due to the difference in maximum intensity. For example, empty frames were characterized as continuous speech. The preset value, on the other hand, was adjusted to 3dB above the background noise which was considered constant. This is a reasonable assumption, based on the fact that activity outside the booth has little effect on the recording (provided that people in the room are reasonably quiet). The DC offset (see section 6.5.3) was kept constant across sessions by keeping the same settings for pre-amplification gain and microphone phantom power.

A comparison between automatic and manual segmentation is shown in Table 9.1. The automatic segmentation yields a 19 % increase in relative duration (Equation 7.5), mostly because of mis-annotation and few exceptional segmentation errors. However, this has little effect on mean pitch and mean intensity calculations, and more significant effect on pitch range and speech rate. The median error shows a less biased estimate of the *expected* error, as the average error is biased by a few extreme cases.

Feature	Cross-correlation	Average error (%)	Median error (%)
Relative Duration	**0.79**	**24.5**	**12.7**
Pitch	**0.82**	**2.8**	**1.7**
Intensity	**0.82**	**1.8**	**1.2**
Pitch range	**0.73**	**16.4**	**12.4**
Speech Rate	**0.74**	**12.9**	**9.4**

Table 9.1: Comparison of manual and automatic segmentation derived a/p feature averages for 43 10-second frames

Automatic segmentation yields consistently larger intervals, but this can be adjusted by means of fine-tuning the intensity and duration thresholds. The trade-off mainly affects the correct segmentation of within-utterance pauses and short utterances competitively. The latter were favored, as they may be the only contributions of the user in a given frame. However, this leads to longer duration and, as a result, to slower speech rate, which is calculated in vowels/minute. Similarly, pitch range is affected by errors in the pitch detection algorithm when applied to non-voiced regions. In conclusion, the overall accuracy was deemed as sufficient for the purposes of the experiment. In order to avoid the effect of extreme errors, the adapted a/p features of the TTS voice

were limited to ±30% of the default voice settings.

Another performance issue identified was that of the *responsiveness* of the system. The experimenter had to type the system utterances in the FreeTTS Player text input box, a process that introduced a delay in the responses of the system. This latency works against the perception of the system as being able to interact in natural dialogue. Whether this would have an effect on the accommodating behaviour of the subjects could be shown only by actually performing the experiments. Nevertheless, some action was taken to remedy the situation, namely that of *pre-loading* the play list of the FreeTTS Player application with a number of common utterances ("yes","no","hello", "do you agree?", etc.) as well as some task-specific words (names of the 15 scenario objects and other scenario-specific words). However, the actual experiments showed that this process did not improve performance significantly, as the experimenter had to type complex responses in order to contribute to the decision process, and there was no way to combine objects in the play list together in order to generate a single utterance. As a result, the experimenter had to type or select shorter utterance fragments, which introduced delays between each fragment. In addition, the TTS synthesizer applied an utterance intonation contour to each fragment, which lead to individual phrases having inappropriate intonation and long utterances to sound "broken".

The TTS voice itself was of very low quality. It is based on diphone *concatenation* (Dutoit 1997), which yields unlimited domain coverage, as any orthographic text can be rendered into speech by combining (concatenating) units (diphones) from a database. The specific TTS voice used was Kevin16, a 16-KHz diphone voice. It is fairly intelligible but sounds robotic and monotonous. FreeTTS also supports MBROLA[26] voices, which are of significantly higher quality. However, implementation of an MBROLA voice in FreeTTS was not possible due to operating system compatibility issues in the available workstations[27].

9.5 Results

The severity of the performance issues described in the previous section became more apparent during the first two actual experiments. These were carried out with a male and a female subject, each participating in two sessions (accommodating vs non-accommodating condition). In particular, the delayed responses of the system voice resulted in the dialogue being significantly slow and "broken". An indication of this is given by the overall JAT which was 0.5 for the male subject and 0.4 for the female subject (in the accommodating conditions). These values are "abnormal", as

26 http://tcts.fpms.ac.be/synthesis/mbrola.html(01/04/2010)
27 FreeTTS does not support MBROLA voices for Microsoft Windows XP and Mac OSX 10.5

typical values for human dialogues are typically higher than 0.7 (see section 8.3.3). Thus, frames of 10 seconds that were used in the online analysis have very short relative duration (less than 0.1) which means that the a/p feature average estimates are not reliable. In addition, subjects tended to solve the task on their own, rarely asking the opinion of the intelligent system. This was mostly the case for the female subject. The male subject had a more cooperative attitude, asking the system for clarification regarding the type and function of objects, but overall both subjects made final decisions on the ranking of the objects on their own. Therefore, the "equality of role" design principle could not be met.

When asked to rate the two systems they had interacted with for "naturalness" on a scale 1-10, both subjects gave equal ratings (male speaker: 6/10, female speaker: 5/10). Therefore, the accommodating behaviour of the system was not perceived explicitly. Importantly, neither subject realized that the system was in fact mediated by a human experimenter, despite the fact that one of the subjects is a speech technology research student. The most likely cause of this is that the speakers perceived the system through an interface metaphor (see section 2.2.4), due to the low quality voice.

The recorded audio files for the accommodating condition underwent off-line analysis, following the procedure described in chapter 7. Due to the sparsity of the utterances, a TAMA frame length of 60 seconds (50% overlap) was used. Significant cross-correlation coefficients were found only for the male speaker for mean pitch and mean intensity. The confidence intervals for the cross-correlograms were large, because the longer frame length leads to less points in the time series (confidence intervals are $\pm 2/\sqrt{N}$, where N is the number of points in each series). A modeling procedure (described in section 7.4.4) yielded the term values shown in Table 9.2. Model A is a model with "fixed" autocorrelation terms, φ_{11} (Equation 7.11) derived from the subject's time series individually, while the feedback terms are estimated by multiple linear regression (see section 7.4.4). Model B is a model in which all three coefficients are estimated by multiple linear regression (Equation 7.10).

Feature	Term	Model A			Model B		
		φ_{11}	φ_{12}	θ_1	φ_{11}	φ_{12}	θ_1
Pitch		0.39	0.41	0.28	0.32	0.42	0.35
Intensity		0.30	0.07	0.41	0.03	0.12	0.56

Table 9.2: Accommodation models for male user interacting with accommodating system

Thus, the large feedback terms φ_{12} and θ_1 indicate convergence from the user towards the system

along these dimensions. The models are not as good fits as those for the human dialogues, thus the evidence of accommodation is not as concrete. Figure 9.3 shows the fitted models plotted along with the actual user values. The models fit well during the first half of the dialogue but less well during the later parts. Exclusion of outlier values such as the last value in each original series might yield a better fit. Interestingly, the zero-lag feedback term, θ_1, is significant for both a/p features, while the lag-one feedback term φ_{12} is significant only for pitch. This would imply a shorter "optimal lag" (Jaffe *et al.* 2001) for accommodation of intensity than for pitch. A physical interpretation of this finding can be that accommodation of loudness (of which intensity is a correlate) occurs more promptly than that of other a/p features. This finding is consistent with the results of (Kousidis *et al.* 2008), where accommodation of intensity was apparent in shorter frame lengths, while accommodation of pitch, pitch range, and speech rate required a longer frame length in order to increase robustness.

The similarity of the two models is, as discussed in section 7.4.4, the result of linear regression, which minimizes the error in one dependent variable based on two (model A) or three (model B) independent variables. Thus, the series are *over-fitted.* The difference between the two models is that model A is biased towards autocorrelation, by keeping the autocorrelation term fixed, while model B finds the best fit based on all three terms. Therefore, the values of the terms are estimates of the contribution of each variable to the minimum error model.

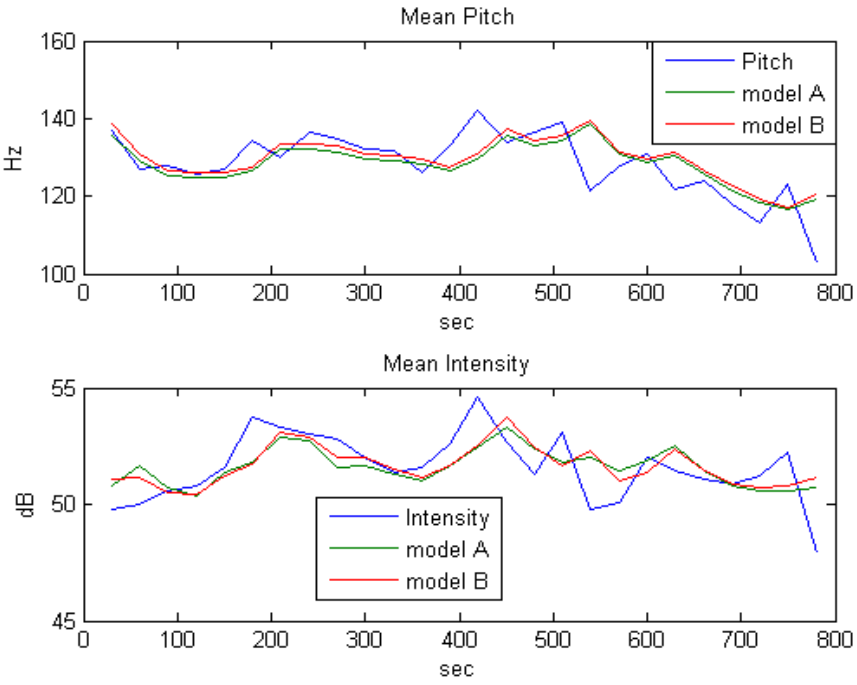


Figure 9.3: TAMA plots of mean Pitch and Intensity and two fitted models (A,B)

9.6 Discussion

This chapter has described a first attempt towards implementing accommodation of a/p features in a simulated (Wizard-of-Oz) SDS environment, based on the TAMA method. As discussed in the previous section, this attempt was generally unsuccessful, due to performance issues which were not resolved due to resource and time constraints. In particular, the main issues were the latency of the system, the robotic, low-quality voice, and the error in the automatic segmentation process, in that order. These issues severely affect the performance of the implementation. The results described in the previous section can be considered as a primary indication of user accommodation (as described by TAMA) at best. This section discusses possible improvements to the experimental design.

The problem of responsiveness, which was the most severe, can be addressed by extending the functionality of the text input interface. One possible extension is to implement an *auto-complete* function, similar to that used in many web interfaces: upon input of one character, the system suggests a list of possible word candidates which makes typing faster. Another improvement could be to implement a touch-screen interface, through which the experimenter can select the suggested words even faster. A further improvement, which was utilized in (Bell *et al.* 2003), would be the addition of another (or more) experimenter(s). In that study, each experimenter was responsible for a different portion of the system's utterances (backchannels, “buying time” for the other experimenter to complete forming an utterance etc). In the experiment described here, this could be implemented as a client-server architecture, in which each experimenter would be able to submit requests through a client interface, while the server would play the utterance queue. It is doubtful whether these improvements would provide for “spontaneous” reaction from the system, but it is likely that the JAT would be increased to a reasonable value (above 0.6), while the “system” would be able to contribute more actively to the task.

The TTS voice problem is easier to rectify, as several alternatives are available freely and commercially. The replacement of Kevin16 with a more natural sounding voice would vastly improve the perception of the system through the *human metaphor* (see section 2.2.4), thus encouraging accommodation from the users that would resemble the behaviour exhibited in human dialogues. It is possible then that the subjects would rate the accommodating system favorably in comparison to a non-accommodating system, but also even more favorably depending on whether they themselves accommodated towards the system. This hypothesis can only be validated *after* an appropriate test platform is implemented.

The implementation of the online prosodic analysis can also be improved. A first improvement would be to implement incremental analysis based on a real-time VAD algorithm. In this case, frames would not be placed at fixed positions, but relative to the generated utterance. Thus, adaptation of a/p features would be based on the immediately preceding context at all times (barring the delay), while still including older frame averages as lag terms. This would enable testing of more elaborate models, which would include lag-zero and lag-one terms, such as the models of equations 7.10 and 7.11. Further, a more accurate VAD algorithm would enable more accurate measurements of the user's a/p features, although the measurement error in the experiments that were carried out was probably the least significant performance issue.

The online analysis and accommodation model component is in itself dialogue and task-independent and can be used in other interaction settings and experiments. It can be implemented in existing SDS architectures, in order to test the perception and evaluation of accommodation in a/p features in existing applications of SDS. In addition, the TAMA methodology is feature independent, which indicates that TAMA-based accommodation models can be used to design systems that accommodate to the user in other modalities, such as head/body movement of the avatar.

As a component of a complete SDS architecture, a TAMA module could share resources with other functions. Such resources include the VAD algorithm, as well as the feature extraction stage (with ASR). Therefore, the addition of accommodating behaviour to existing SDS by means of a TAMA-based module would add a negligible amount of computational load, namely the VAR model equations. The best place to add the TAMA module in the general SDS architecture (see Figure 2.2) is the interaction manager, as conceptually inter-speaker accommodation is a behaviour related to the interaction between interlocutors. However, prosodic adaptation would have to be implemented in the utterance generation phase, as a modification of the input to the TTS module. In case of a system with pre-recorded prompts, prosodic modification could be performed online on the prompts, or, in case this is computationally expensive, several instances of the prompts with different prosodic characteristics could exist in the prompt database.

In conclusion, while the experiment described in this chapter failed due to technical limitations and design inadequacies, the suggested improvements point to an appropriate implementation of accommodation in SDS that can test user perception and user response to such behaviour in accordance to the human metaphor paradigm.

10 Conclusions and future work

10.1 Conclusions

The main objective of this thesis, as stated in the introduction, was the the formulation of a quantitative description of inter-speaker accommodation of prosodic and temporal features in spontaneous human dialogues that is useful from the point of view of SDS, in view of implementing similar behaviour in human-machine interaction where appropriate.

The above objective was pursued by the formulation of TAMA and the statistical modeling of accommodation that was presented in chapter 7. This approach proved sufficient for a/p features, as the feature averages calculated from overlapping frames were robust. The analysis revealed a picture of ubiquitous accommodation for mean pitch and mean intensity, while accommodation of speech rate and pitch range was less common. However, the measurements for the latter features were less robust and an optimization of the automatic feature extraction procedure may yield results which are comparable to those found for mean pitch and mean intensity.

The statistical models presented in section 7.4.4 provide a measurement of the strength, or degree of accommodation for each speaker, namely the feedback terms of the models. However, these models assume accommodation as deterministic and other variations as random, thus the actual coefficient values are valid for across-speaker comparison purposes only: they indicate whether speakers accommodate their features or not, the direction of accommodation (uni-directional or bi-directional), as well as which speaker accommodates more towards the other. They do not indicate the portion of variation in a/p features that is accounted for by accommodation. The latter can be estimated by the strength of regression (R^2), but this estimate is biased unless other sources of variation are taken into account and a principal component analysis is perfomed.

An application of TAMA to temporal features was presented in chapter 8. TAMA analysis of mean switch-pause duration and frequency of overlap speech during turn exchanges was not as powerful as in the case of a/p features, due to data sparsity. As a result, there is a more severe trade-off between resolution of the representation and reliability of the calculated frame feature averages. Further, TAMA analysis of temporal features resulted in similar conclusions with those of (Edlund *et al.* 2009): only a portion of the analyzed dialogue shows synchronous accommodation of these features. Similarly to (Edlund *et al.* 2009), it was concluded that additional sources of variation "override" accommodation of these temporal features.

A novel dialogue representation, comprising turn shares and turn share distributions was presented in chapter 8. This representation can be derived directly from a chronograph of the dialogue and approaches the problem of turn-taking from a different angle, namely disregarding the idea of turns

and considering both speakers as simultaneously active. Turn shares express the proportional amount of contribution of each interlocutor in a given time-frame, while turn distributions express the overall activity, in terms of the proportional amount of vocalization,overlap and silence in the dialogue. The proposed representation proved useful in accounting for a significant amount of variation in average switch-pause length and frequency of interrupting overlaps, as they were found to be correlated to dialogue activity (as expressed by JAT), turn share and exchange rate (ER), which is derived from turn share. In addition, a follow-up analysis based on the proposed representation provided evidence towards attributing the across-dialogue linear relationship of average pause length across speakers, found in (Bosch *et al.* 2005), to inter-speaker accommodation, rather than to dialogue liveliness. Although average pause length is correlated to liveliness, it was shown that this correlation does not account for the similarity across speakers, although there was some evidence of similarity increasing with liveliness. The latter finding was also reported in (Nishimura *et al.* 2008).

Although a model was not formulated for temporal features, as in the case of prosodic features, the findings provide useful insights for SDS with conversational capabilities (i.e. not half-duplex turn-taking which is the current norm). One of these insights is the implementation of the turn-share representation in SDS interaction management, as it could be used to improve end-pointing by monitoring the turn-share distribution online: in an end-pointing approach such as that of (Raux 2008), which uses silence thresholds according to detection of TRPs based on the dialogue context (prosodic or semantic), online adaptation of the thresholds according to the turn-distribution could improve performance. In contrast, synchronous accommodation of switch-pause length would not be a sufficient strategy for "free-talk" SDS, as this model is too simplistic to characterize human communication. It is possible, however, that such a model would be sufficient for applications that are half-duplex by definition (e.g. information retrieval or travel booking), in which case there is a straightforward succession of turns in the interaction.

A preliminary experiment of implementation of an a/p accommodation model in an SDS environment was presented in chapter 9. Several performance issues hindered the possibility of acquiring useful information from these experiments. However, some components of the method, such as the online prosodic analysis and monitoring module performed well and could be re-used in other experiments, while the testing platform can be significantly improved by a number of optimizations described in chapter 9. In addition, in one of the two experiments performed, the user was found to moderately accommodate his pitch and intensity to that of the system. Therefore, there is at least a minor indication of continuous user accommodation towards a TTS voice, which is in

agreement with findings of studies that performed comparisons across dialogues (Oviatt *et al.* 2004; Suzuki and Katagiri 2005). An optimization of the experimental design is required in order to evaluate the magnitude of benefits for SDS.

10.2 Future work

A number of possible extensions of the work presented in this dissertation have already been discussed in chapters 7,8 and 9. This section summarizes some of these directions for future work.

(a) Extensions to TAMA: A combination of TAMA with dialogue act categorization, which can be obtained by one of the automatic classification methods presented in section 2.4.4, would yield a more accurate description of accommodation of a/p features. Such an approach would comprise a global feature mean per dialogue act type, and individual utterance features would be normalized over their respective global mean, prior to calculating the frame average. Thus, variation due to the inherent prosodic properties of dialogue acts would be accounted for prior to assessing accommodation. This approach is appealing because dialogue act classification can be performed automatically, using prosody as a classifier. Another possible extension of the work presented here is the application of TAMA to other modalities (body/head movement) or to other measurements of the same features (e.g. pitch/intensity of stressed syllables), which is relatively straightforward as TAMA is feature-independent.

(b) Extensions to the statistical model: The bi-variate models presented in section 7.4.4 consider variation due to accommodation as deterministic, while the random component accounts for all other variation (utterance-specific, paralinguistic, emotional content). The models could be optimized by including exogenous factors that separately model other sources of variation (e.g. an emotional model). Another possibility is to formulate a model which considers co-integration (see section 7.4.4). This approach would comprise a linear combination of the two component series, such as their absolute difference, which can be considered as a measure of distance between interlocutors (for normalized feature averages). Finally, accommodation along different modalities can be modeled simultaneously, in an n-variate model, which would include different features from each speakers as independent variables. However, such models are characterized by increased complexity: the increased number of independent variables significantly increases the possibility of a type I error, i.e. cross-correlations could in fact occur randomly. This disadvantage is counter-balanced by the possibility to assess accommodation of different features that may have the same underlying cause: for example, the effort code (Gussenhoven 2005), is manifested in both pitch range and

articulation precision. Thus accommodation of effort could be manifested along different modalities by each interactant.

(c) Extensions to studying temporal accommodation: Variations in pause duration and frequency of overlaps can be attributed to many factors. Perhaps a possible route is the combination of discourse analysis and accommodation measurements, such as the application of TAMA presented in chapter 8.2 and the approach of (Raux 2008). In this case, actual pause durations could be normalized (z-scored) according to a normal distribution of durations following specific dialogue acts. Similarly, the occurrence of an overlap could be z-scored according to the probability of an overlap occurring at a specific point in the discourse. This would enable an assessment of accommodation based on whether interactants tend to shorten their silent intervals synchronously, while taking variations that are dialogue-act specific into account. Another possible route is to investigate the relationship between temporal organization and speech rate, either by accounting for pause duration shortening due to faster delivery rate, or by investigating whether accommodation of speech rate and temporal features tend to co-occur. Finally, it is possible to combine the turn share representation presented in chapter 8 with a serial approach, such as TAMA or the one proposed in (Edlund *et al.* 2009), in order to compare temporal features across speakers, while accounting for variation due to turn share and liveliness, as expressed by JAT.

(d) Extensions to the SDS implementation test platform: In the absence of SDS that can engage in human-like conversation, Wizard-of-Oz experiments are the most plausible solution of evaluating the benefits of implementing accommodation in human-machine interaction. The performance optimizations described in section 9.6 (faster prompt generation/selection interface, multiple experimenters, more accurate VAD algorithm, better TTS voice) can provide for an adequately human-like conversation, in order to investigate continuous accommodation of a/p features in human-machine dialogues. Another option is to include the online prosodic analysis module and accommodating model in existing SDS architectures and applications, for which subjective evaluation procedures are well-established (Moller *et al.* 2007). An analogous evaluation approach for temporal features is far more challenging, as any kind of utterance generation, signal manipulation or decision process invariably introduces latencies before system prompts, thus making accommodation of temporal features difficult. However, it is still possible to assess the benefits of temporal accommodation by considering micro-domains (Edlund *et al.* 2009), in which the interaction is so constrained that latencies can be minimized.

APPENDIX A: Recorded dialogues and analysis results

This appendix presents additional information on the corpus of dialogues acquired as described in section 6.4. The dialogues are categorized in three types:

a) Dialogues recorded using the “shipwrecked” scenario process (section 6.4.3). These dialogues are coded 's*n*' in the tables below, and they include the additional two scenarios “space pod” (Figure A.1) and “Himalayas” (Figure A.2). The specific scenario is denoted in a separate column, named 'info' in the tables.

b) Dialogues using a MIP procedure, in which participants are given a score every time they rank one of the items (section 6.4.3). These dialogues are coded 'ms*n*', and the specific scenario used in the session is denoted in the 'info' column.

c) Dialogues which comprise unconstrained conversation between two participants situated in the isolation booths (section 6.4.1). These dialogues are coded 'u*n*' in the tables. The 'info' column contains the main topic of conversation adopted by the speakers.

FIGURE A.1 – Himalayas scenario

FIGURE A.2 Space pod scenario

TABLE A.1 – Prosodic Features

Dialogue	Info	SPEAKER A					SPEAKER B				
		Gender	Mean Pitch (Hz)	Mean Intensity (dB)	Pitch Range* (Hz)	Speech Rate (vowels/ min)	Gender	Mean Pitch (Hz)	Mean Intensity (dB)	Pitch Range* (Hz)	Speech Rate (vowels/ min)
s1	Shipwrecked	M	121	70.5	45	227	M	125	73.3	44	200
s2	Shipwrecked	M	113	65.6	41	250	F	202	60.2	104	246
s3	Himalayas	M	139	72.1	31	214	M	109	66.1	26	227
s4	Nuclear	M	105	72.0	23	236	M	119	70.8	37	257
s5	Shipwrecked	F	215	70.8	96	194	M	164	70.0	30	233
s6	Shipwrecked	M	103	68.3	24	231	M	161	61.7	30	222
s7	Shipwrecked	M	136	69.0	42	224	M	163	69.3	39	229
s8	Shipwrecked	F	199	63.6	110	239	F	210	60.4	79	226
s9	Space pod	F	193	60.2	75	242	M	137	66.7	44	173
s10	Space pod	M	117	69.4	25	227	M	144	69.1	46	178
s11	Shipwrecked	M	144	79.7	32	216	M	138	62.0	33	217
s12	Space pod	M	142	62.4	51	210	F	222	61.2	121	192
s13	Himalayas	F	197	61.6	71	221	M	130	65.2	38	175
s14	Shipwrecked	M	140	70.7	41	183	M	166	59.8	39	216
ms1	Shipwrecked	F	225	57.4	93	228	F	214	62.4	96	225
ms2	Shipwrecked	F	253	69.7	136	245	M	127	70.3	51	209
ms3	Shipwrecked	F	199	67.3	68	243	F	182	65.7	95	262
ms4	Shipwrecked	F	228	73.3	80	225	F	237	67.6	100	220

ms5	Shipwrecked	M	139	68.1	42	193	M	155	62.8	42	211
ms6	Shipwrecked	M	118	72.8	44	173	M	121	70.3	38	219
ms7	Shipwrecked	F	208	60.4	109	229	F	196	63.2	75	220
ms8	Shipwrecked	M	123	64.5	81	216	M	122	69.9	37	186
u1	Entertainment	M	128	77.5	36	201	M	135	77.1	53	161
u2	Sports/Scotland	M	140	68.0	41.3	230	M	118	66.2	38	222
u3	Various	M	121	76.1	50	190	M	130	74.6	52	224
u4	Children	M	118	74.5	59	191	M	113	76.9	46	167
u5	Environment	M	131	65.7	54	164	M	120	61.2	50	214
u6	Work, society	M	107	47.0	90**	213	M	126	48.0	127**	236
u7	Work in Ireland	M	128	71.0	74**	221	M	125	73.0	73**	243
u8	Study, Slovenia	M	110	60.0	79**	217	F	167	61.0	132**	289

* Pitch range calculated as $2*p_{std}$, the standard deviation of speech interval pitch
** Pitch range calculates as $p_{max} - p_{min}$, the maximum and minimum pitch values in the speech interval pitch contour (not stylized)

TABLE A.2 – Turn Distribution and duration

dialogue	TA (%)	TB (%)	TP (%)	TO (%)	TSA	TSB	JAT	TDD (sec)
s1	44.9	29.7	19.7	5.7	0.59	0.41	0.80	524
s2	43.8	14.6	36.7	4.9	0.71	0.29	0.63	622
s3	37.5	33.5	20.7	8.3	0.52	0.48	0.79	488
s4	50.0	25.1	18.3	6.6	0.64	0.36	0.82	280
s5	27.2	28.1	40.2	4.5	0.49	0.51	0.60	465
s6	33.0	23.1	37.1	6.9	0.57	0.43	0.63	351
s7	35.9	28.3	31.9	3.9	0.52	0.48	0.68	636
s8	29.0	22.1	38.9	10.0	0.55	0.45	0.61	517
s9	20.5	43.9	18.4	17.2	0.38	0.62	0.82	428
s10	35.6	32.9	16.5	15.1	0.51	0.49	0.84	492
s11	49.3	18.2	27.3	5.3	0.70	0.30	0.73	595
s12	40.3	24.3	22.1	13.3	0.59	0.41	0.78	384
s13	18.3	46.4	20.2	15.1	0.35	0.65	0.80	354
s14	42.4	29.8	20.9	6.9	0.57	0.43	0.79	354
ms1	19.5	28.9	45.5	6.1	0.42	0.58	0.55	584
ms2	25.3	32.6	31.0	11.1	0.45	0.55	0.69	201
ms3	32.2	30.1	30.1	7.6	0.51	0.49	0.70	643
ms4	34.2	19.2	37.6	9.0	0.61	0.39	0.62	610
ms5	40.8	15.0	42.7	1.5	0.72	0.28	0.57	614
ms6	31.3	27.0	28.3	13.4	0.53	0.47	0.72	683
ms7	23.8	30.0	37.0	9.2	0.46	0.54	0.63	599

ms8	34.5	25.7	32.1	7.7	0.56	0.44	0.68	595
u1	49.9	27.8	8.6	13.7	0.61	0.39	0.91	1728
u2	35.1	45.3	8.5	11.1	0.45	0.55	0.92	538
u3	39.1	38.1	6.1	16.7	0.50	0.50	0.94	813
u4	32.5	51.1	6.5	9.9	0.41	0.59	0.94	403
u5	40.0	37.1	15.1	7.8	0.52	0.48	0.85	1266
u6*	-	-	-	-	-	-	-	1781
u7*	-	-	-	-	-	-	-	1805
u8*	-	-	-	-	-	-	-	1330

* Dialogues analyzed in (Kousidis *et al.* 2008)

TA: Percentage of vocalization by speaker A
TB: Percentage of vocalization by speaker B
TP: Percentage of silence
TO: Percentage of joined (overlapping) vocalization
TSA: Turn share of speaker A
TSB: Turn share of speaker B
JAT: Joint active time
TDD: Total dialogue duration

TABLE A.3 – TAMA of A/P features and statistical evaluation of accommodation

Dialogue	Frame length: 30 sec Time step: 20 sec				Frame length: 20 sec Time step: 10 sec			
	Pitch	**Intensity**	**Pitch range**	**Speech rate**	**Pitch**	**Intensity**	**Pitch range**	**Speech rate**
s1	-	-	-	0	-	-	-	-
s2	-	-1	-	-	-	-	1	-
s3	-	-	0(-)	1*	-1(-)	-	-	-
s4	-	-	-	-	1	0	-	-
s5	-	-	-	-	0	0,1*	-	-
s6	-	-1	0	-	-	-	-	-
s7	-	0	-	1	1	0	-	-1,1
s8	-	-	-	-	0	0	0	-
s9	-	-	-	-	0,1	0,1	1	-1
s10	-	-	-	-	0	0	0	-
s11	-	-	-	-	-	0	0	-
s12	-	-	-	-	1	0	-	1
s13	-	-	-	-	0	0	0*	-
s14	-	-	-	0*	0,1	-	-	-
ms1	-	-	-	-	0,-1	0,-1	0	-
ms2	-	-	-	-	-	-	1*	0*
ms3	-	-	-	-	0	0	-	0
ms4	-	-	-	-	-1,0*	0	0	-
ms5	-	-	0*	1*	0	0	-	-
ms6	-	-	-	-	0	0	0	1
ms7	-	-	0	-	0	0*	-	-
ms8	-	-	0	-	0	0*	-	-
u1	1	-1	1	-	-1*	-	-	-
u2	0	-	-	1*	0*	0*	1	-
u3	-	0	0*	-	0,1	0*	-	0*
u5	-	-	-	0	-	0	1	-

* cross-correlation coefficients significant at 90% confidence intervals, all other coefficients significant at 95% confidence intervals
Numbers indicate lags at which positive coefficients are found in the cross-correlogram
(-) signifies a negative coefficient

TABLE A.3 – Average pause length and overlap rate

Speaker	A						B					
Dialogue	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**
s1	451	344	414	330	331	18,5	493	298	493	338	347	30,6
s2	710	374	532	513	483	12,3	544	334	442	350	374	26,9
s3	413	350	413	336	313	32	431	301	431	315	293	28,8
s4	343	307	319	256	267	24,4	375	324	375	312	290	37,7
s5	907	293	492	420	436	18,9	865	541	395	541	541	22,3
s6	890	282	381	264	333	23,3	412	300	384	265	273	35,3
s7	687	300	487	421	410	14,8	682	299	447	424	410	23,1
s8	775	296	644	535	502	30,4	1017	309	544	632	597	28,2
s9	335	290	335	224	222	44	353	310	353	296	268	41,2
s10	382	359	382	299	296	42,3	412	326	366	288	282	45
s11	728	331	547	477	509	18,9	505	234	424	259	278	34,5
s12	483	345	483	369	345	35,9	495	326	495	344	364	34,5
s13	339	317	339	262	264	48,6	408	284	408	300	288	35
s14	492	359	457	382	345	20,5	361	319	361	297	243	32,7
ms1	921	303	738	847	685	39,8	1021	304	729	821	760	15,2
ms2	662	401	541	511	454	25	926	299	567	460	515	33,3
ms3	680	274	589	436	449	31,5	608	362	456	382	382	23,3
ms4	845	321	516	505	513	30,1	844	282	595	532	499	34,5
ms5	926	315	654	680	634	6,6	834	335	581	560	580	12,8

ms6	590	330	540	415	427	34,7	479	367	464	368	358	30,1
ms7	862	277	683	584	514	30,7	884	322	543	564	548	31,6
ms8	696	317	599	482	451	25,7	673	257	482	404	380	36,8
u1	296	292	296	232	234	37,3	405	365	405	328	276	68,1
u2	386	331	386	349	277	42,6	543	396	437	430	385	39,1
u3	288	246	288	212	199	62,1	371	341	371	320	281	63,8
u4	405	380	405	352	306	57,5	281	281	281	224	228	45,1
u5	448	377	435	360	354	39,9	470	371	461	360	341	22,7

APL: Average pause length (arithmetic mean of original pause duration distribution)
d<1: Distribution skewness corrected by applying a duration threshold of 1 seconds
d<2: Distribution skewness corrected by applying a duration threshold of 2 seconds
APL median: Distribution skewness corrected by taking median value instead of arithmetic mean
APL log: Distribution skewness corrected by taking the arithmetic mean of a log transformed distribution (mean transformed back to ms)
OR: Overlap rate (percentage of speaker utterances initiated during partner vocalization)

TABLE A.4 – Average switch pause and interrupting overlap rate

Speaker	A						B					
Dialogue	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**
s1	408	327	350	288	254	27,3	664	279	664	476	422	40,9
s2	676	306	466	401	361	24,7	656	375	460	496	440	31,9
s3	387	317	387	336	277	50,6	406	305	406	320	254	28,8
s4	233	233	233	208	176	38,6	324	324	324	268	251	37,8
s5	1164	283	464	460	421	26,9	799	421	567	564	530	19,4
s6	1159	298	331	282	324	35,1	378	263	378	267	230	32,1
s7	585	313	477	403	333	25,5	750	264	458	418	416	25,3
s8	723	288	723	557	476	31	957	322	459	592	532	30,6
s9	303	261	303	197	179	46,6	248	248	248	241	201	48,6
s10	288	288	288	213	207	50,5	348	299	348	256	235	40,6
s11	795	311	372	405	468	41,1	470	223	470	256	242	37,5
s12	425	326	425	293	275	36,1	483	323	483	356	348	42,3
s13	375	342	375	300	281	48,5	279	240	279	238	189	58,8
s14	442	289	365	286	261	31,3	349	290	349	276	210	27,9
ms1	668	310	668	548	477	54,4	1148	292	767	960	852	12,8
ms2	683	371	537	552	365	35,8	1114	293	718	664	630	28,6
ms3	512	242	512	320	333	37	536	337	455	320	341	23,5
ms4	779	299	478	453	433	37,3	802	267	544	506	452	34,4
ms5	783	293	652	616	493	7,5	711	347	578	473	544	7,5
ms6	531	310	484	373	367	36,9	434	332	434	344	316	29,7

ms7	815	263	584	503	390	35,6	862	265	554	564	433	39,1
ms8	646	304	612	520	382	29,1	425	231	392	249	237	46,2
u1	179	179	179	152	125	64,9	371	323	371	256	231	50,8
u2	310	227	310	206	201	39,5	396	340	396	269	285	51,3
u3	244	213	244	176	167	51,9	286	255	286	192	191	64,4
u4	266	266	266	176	131	62,5	150	150	150	120	129	56
u5	382	339	382	320	275	44,4	259	226	259	193	163	34,6

APL: Average switch pause length (arithmetic mean of original switch pause duration distribution)
d<1: Distribution skewness corrected by applying a duration threshold of 1 seconds
d<2: Distribution skewness corrected by applying a duration threshold of 2 seconds
APL median: Distribution skewness corrected by taking median value instead of arithmetic mean
APL log: Distribution skewness corrected by taking the arithmetic mean of a log transformed distribution (mean transformed back to ms)
OR: Overlap rate (percentage of speaker turns initiated during partner vocalization)

TABLE A.5 – Switch pause and overlap rate cross-correlation (60/30)

Frame length: 60 Time step: 30						
Dialogue	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**
s1	0,44*	-	-	-	-	-
s2	-	-	-	-	-	-
s3	-	0,41*	-	-	-	-
s4	-	-	-	0,74	0,78	-
s5	-	-	-	-	-	-
s6	-	-	-	-	-	0,7
s7	-0,47	-	-	-0,41	-0,49	0,58
s8	-	-	-	-	-	-
s9	-	-	-	-	-	0,46
s10	-	-	-	-	-	-
s11	-	-	-	-	-	-0,39
s12	-	-	-	-	-	0,78
s13	0,6	-	0,6	0,51*	0,54*	-
s14	0,71	-	-	-	-	0,65
ms1	-	-	-	-	-	-0,52*
ms2	-	-	-	-	-	-
ms3	-	-	-	-	-	-
ms4	0,73	-	0,43*	0,64	0,64	-
ms5	-	-	-	-	-	-
ms6	-	-	-	-0,69	-	-
ms7	-	-0,6	-	-	-	-
ms8	-	-	-	-	-	-
u1	-	-	-	-	-	-
u2	-	-	-	-	-	-
u3	-	-	-	-	0,42	-
u4	-	-	-	-	-	-
u5	-	-	-	-0,31	-	-

* $p<0.10$ all other coefficients significant at $p<0.05$

TABLE A.6 – Switch pause and overlap rate cross-correlation (30/20)

Frame length: 30 Time step: 20						
Dialogue	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**
s1	0,37*	-	-	-	-	0,66
s2	-	0,36*	-	-	-	-
s3	-	-	-	-	-	-
s4	-	-	-	-	-	0,48*
s5	-	-	-	-	-	-
s6	-	-	-	-	-	0,66
s7	-	-0,32	-0,34*	-	-	-
s8	-	-	-	-	-	-
s9	-	-	-	-	-	0,6
s10	-	-	-	-	-	-
s11	-	-	-	-	-	-
s12	-	-	-	-	-	0,71
s13	-	-	-	-	-	-
s14	0,82	0,59	-	0,78	0,6	-
ms1	-	-	-	-	-	-0,72
ms2	-	-	-	-	-	-
ms3	-	-	-	-	-	-
ms4	0,33*	-	-	0,44	0,30*	-
ms5	-	-	-	-0,3	-	-
ms6	-	-	-	-	-	-
ms7	-	-	-	-	-	-
ms8	-	-	-	-	-	-
u1	-	-	-	-	-	-
u2	-	-	-	-	-	-
u3	-	-	-	-	-	-
u4	0,83	0,83	0,83	0,83	0,85	-
u5	-	-	-	-	-	0,41

* $p<0.10$ all other coefficients significant at $p<0.05$

TABLE A.7 – Switch pause and overlap rate cross-correlation (20/10)

Frame length: 20 Time step: 10						
Dialogue	**APL (ms)**	**APL d<1 (ms)**	**APL d<2 (ms)**	**APL median (ms)**	**APL log (ms)**	**OR (%)**
s1	-	-	-	0,36	0,38	0,63
s2	-	-	-	-	-	-
s3	-	-	-	-	-	-
s4	-	-	-	-	-	-
s5	-	-	-	-	-	-
s6	-	-	-	0,31*	-	0,78
s7	-	-	-	-	-	-
s8	-	-	-	-	-	-
s9	-	-	-	-	-0,29*	
s10	-	-	-	-	-	0,33*
s11	-	-	-	-	-	-
s12	-	-	-	-	-	0,56
s13	-	-	-	-	-	-
s14	0,7	-	-	0,6	0,41	
ms1	-	-	-	-	-	-0,61
ms2	-	-	-	-	-	-
ms3	-	-	-	-	-	-0,31*
ms4	-	-	-	-	-	-
ms5	-	-	-	-	-	-
ms6	-	-	-	-	-	-
ms7	-	-	-	-	-	-
ms8	-	-	-	-	-	-
u1	-0,26	-	-	-0,23*	-0,24*	0,18*
u2	-	-	-	-	-	-
u3		0,30*	-	-	-	0,27
u4	-	-	-	-	-	-
u5	-	-	-	-	-	0,49

* p<0.10 all other coefficients significant at p<0.05

APPENDIX B: Vowel detection and speech rate estimation

This appendix presents the results of a performance test which was carried out in order to evaluate the accuracy of four different syllable/vowel detection methods for the purpose of speech rate estimation (see section 6.5.3). All four methods were implemented as Praat scripts.

The criteria of selection for these methods were (a) ASR independency, (b) language independency, (c) non-requirement of training data, enabling online estimation, and (d) low computational cost. Thus the following four methods were selected for testing:

Method 1: Modified beat extractor

(Cummins and Port 1998) presented a "beat extractor" software, with beats being "very close to" vowel onsets. The process comprised a filter bank of 6 gammatone filters in the range 300 to 2000 Hz. The purpose of accentuating energy in this region is to amplify the effects of the first two formants, while F0 and high-frequency energy - which is typically the result of frication – are mostly filtered out. The six energy contours are smoothed and summed, yielding a single energy measure which is low-passed filtered (20 Hz). (Cummins and Port 1998) defined beats as the medium points between dips and peaks in this smoothed energy contour, and noted that they occur very closely to vowel onsets.

Since the purpose of the definition of beats was not to enumerate them, (Cummins and Port 1998) did not report performance test results. (Barbosa 2009) reported using a modified version of this method for vowel onset detection in combination with manual corrections, and also did not report performance test results of detection accuracy.

The method implemented here is an unpublished modification of the beat extractor, implemented as a Praat script. This script was developed by Hugo Quené and is available online[28]. It uses the derivative of the intensity contour to identify steep rises in the intensity contour of a filtered speech signal, which typically coincide with vowel onsets. The filter is a pass-band in the range 500-1000 Hz. Steep rises in intensity are identified as local maxima in the derivative of intensity contour. The vowel onsets are assumed to be "half-way" between the local maximum and the moment of maximum intensity. This method has been used for vowel detection in LinguaTag (Cullen 2008b), a multipurpose speech annotation tool developed in the SALERO project.

28 http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/vowelonset_v3_praat.txt

Method 2: Syllable detection based on intensity contour

The syllable detection proposed in (deJong and Wempe 2007), implemented as a Praat script which is available online[29]. Potential syllables are detected as peaks in the speech signal's intensity contour, with a peak threshold set at 0 or 2 dB over the median intensity of the signal (depending on whether the signal is pre-filtered). The preceding "dip" of the intensity contour, prior to the peak, is considered in order to discard peaks that are not 2 or 4 dB "louder" than the preceding dip (again depending on pre-filtering). In a third step, peaks that are located outside voiced regions, are discarded. All remaining peaks are considered as syllables and are annotated with a single boundary at the peak location. Using this method, (deJong and Wempe 2007) achieved a correlation of 0.7 between automatically detected and hand-labeled syllable rates in speech "spurts" with a fixed length of 5 seconds. Correlation was higher (0.8-0.88) for entire speech files. (deJong and Wempe 2007) noted that the detection algorithm misses mostly unstressed syllables. The comparison of this method to the other 3 is based on the assumption of a 1:1 syllable/vowel ratio.

Method 3: Derivative of intensity in F1-F2 frequency band

This method is reported in (Barbosa 2009) and is implemented as a Praat script which was kindly provided by the author of the study, Plinio Barbosa. This method uses the derivative of the smoothed energy contour of the beat extractor (Cummins and Port 1998) in order to locate steep rises in the energy of the F1-F2 frequency band. The method requires a setting for speaker gender (male or female).

Method 4: Original beat extractor

The original beat extractor (Cummins and Port 1998), provided as an option is the script provided by Plinio Barbosa (see method 3 above).

Test corpus

The test corpus was a collection of speech intervals, taken from one of the dialogues in the "shipwrecked" corpus (s2 – see appendix A). The dialogue is between a male and a female participant. Approximately half of this dialogue was used for this comparison test. Speech intervals were automatically segmented and manually corrected following the method described in section 6.5.1. The speech intervals were manually transcribed and the number of syllables in each

29 http://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei

transcribed sentence was counted. Intervals which comprised nonsense words, laughing speech or voiceless speech (whisper) were excluded. The results of the manual vowel detection are shown in Table B.1 below:

TABLE B.1 – Reference syllables from the "shipwrecked" corpus

Speaker	Total number of speech intervals	Total duration (sec)	Total number of vowels
Male	62	63.43	349
Female	45	43.62	188

Vowels or syllables (in the case of method 4) were automatically detected using each of the four detection methods. The results are shown in Table B.2. It is evident that all methods miss a significant number of vowels. The most poorly performing method in that regard is method 2, while the best performance is obtained using method 1 (LinguaTag).

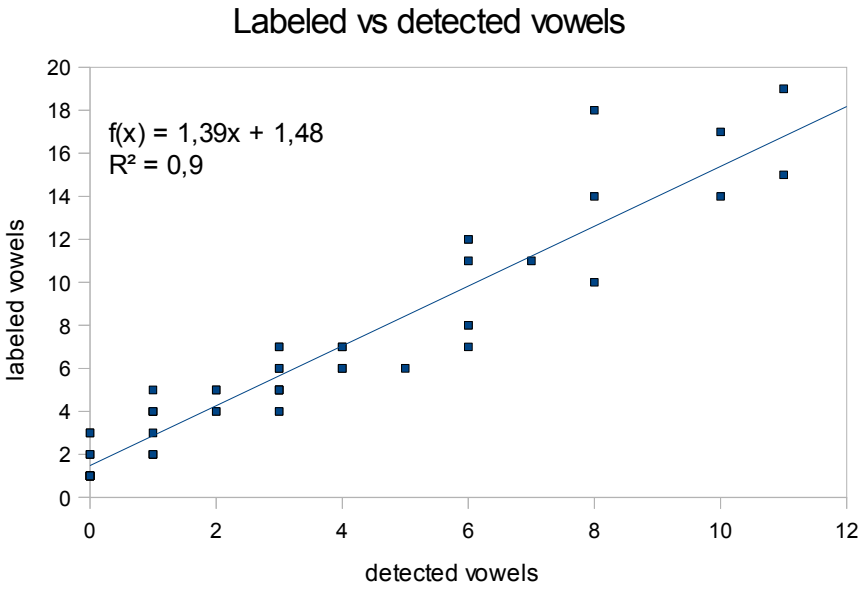
TABLE B.2 – Comparison performance of the four automatic syllable/vowel detection methods

Speaker	Method	Number of detected vowels	Error (%)	Correlation
Male	**ref**	349	-	-
	1	258	26	0.89
	2	185	47	0.95
	3	250	28	0.86
	4	241	31	0.92
Female	**ref**	188	-	-
	1	169	10	0.95
	2	96	49	0.95
	3	150	20	0.74
	4	152	19	0.92

However, method 2 shows the best correlation to the reference syllable count for both male and female speakers: it consistently detects approximately half of the manually labeled syllables and can therefore be used as an estimate of speech rate. The actual number of vowels can be calculated by

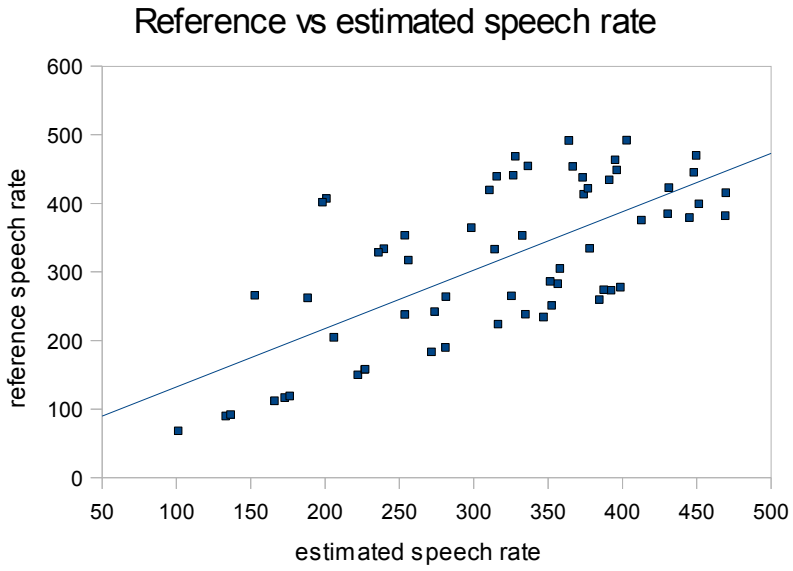
the detected vowels using a linear regression model, as shown in Figure B.1 below:

FIGURE B.1 – Scatter plot of labeled vs detected vowels



An estimate of speech rate is then calculated using the value derived from the regression model. The correlation of the estimate with the reference measure (derived from the hand-labeled vowels) is equal to 0.69.

FIGURE B.2 – Speech rate estimation using syllable detection and linear regression



Method 3 has the poorest correlation to the reference vowel count and is thus the least suitable

method for speech rate estimation. Finally, method 4 is performing quite well both in number of vowels detected as well as correlation to the reference count. Both this method and method 1 (LinguaTag) are based on taking the derivative of the energy contour in order to identify steep rises adjacent to peaks. Their performance can be further optimized by adjusting their threshold parameters.

In conclusion, the vowel detection method used in LinguaTag (method 1) has the best performance (smallest % error of detected vowels), while yielding a high correlation to the reference vowel count, which also makes it a good estimate of speech rate, with results comparable to those of (deJong and Wempe 2007) (method 2). Thus, this method was chosen for speech rate estimation in the feature extraction phase described in section 6.5.3.

APPENDIX C: Code implementations

This appendix provides the listings of several Praat and MATLAB scripts and commands which implement the feature extraction and analysis procedures described in this dissertation. The presentation order follows a step-by-step analysis of a recorded dialogue using Praat and MATLAB software.

1. Segmentation and annotation

As mentioned in section 6.5.1, automatic speech/silence segmentation is a built-in command in recent versions of Praat:

```
annotate → to textgrid(silences)
```

The command is available on the dynamic menu only if a Sound object is selected. The input parameters are: (1) the *silence threshold* (in dB relative to the maximum intensity and therefore always negative), (2) the *minimum silent interval duration* (in seconds), the *minimum sounding interval duration* (in seconds), (3) the *silent interval marker* and the *sounding interval marker* (preferred strings), (4) the *minimum pitch* (in Hz, which is required for the signal intensity calculation) and (5) the *time step* (in seconds, which can be used to adjust the resolution of the intensity analysis). The output of this command is a TextGrid object which contains boundaries of sounding/silent regions of the signal. This object can be edited in combination with the sound file (Sound object) in the Praat environment in order to perform manual corrections to the segmentation and set appropriate labels to the non-speech intervals (see section 6.5.2).

However, the definition of the silence threshold relative to the maximum is problematic for long sound files (such as entire dialogues), since a single global maximum is not the best reference value for all the various portions of the dialogue with varying intensity. In addition, if the sound object contains only silence, the command recognizes it as speech (intensity is roughly constant near the maximum throughout). This especially problematic for online analysis of TAMA frames as described in section 9.3, where it is possible that frames are void of speech.

The following script (Listing C.1) overcomes these problems by implementing the same method as the built-in command, but defining the intensity threshold as an absolute value (in dB). This should be set at 1-3 dB over the background noise intensity. All other parameters and output are the same as in the built-in command.

Listing C.1: Speech/silence segmentation script

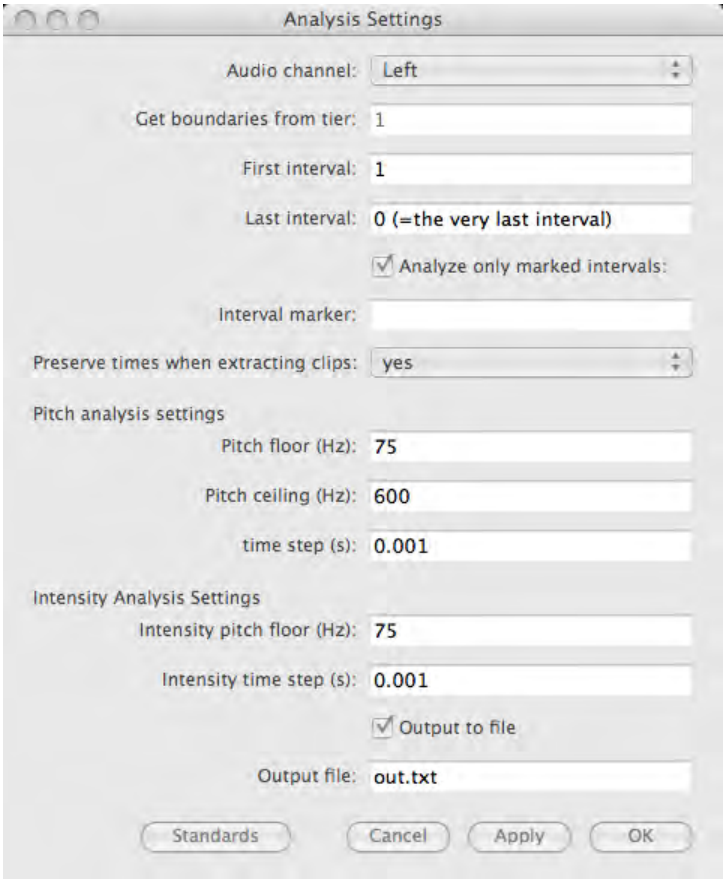
```
#segmentation of signal into speech/silence
form Annotate Silences
      real minimum_pitch_(Hz) 75
      real time_step_(s) 0.0 (=auto)
      real Background_noise_threshold_(dB) 35
      real Minimum_sounding_duration_(s) 0.25
      real Minimum_silent_duration_(s) 0.1
      sentence sounding_marker s
      sentence silent_marker p
endform

minpitch= minimum_pitch
tstep = time_step
nthresh = background_noise_threshold
minsound = minimum_sounding_duration
minpause = minimum_silent_duration
smark$ = sounding_marker$
pmark$ = silent_marker$

sound = selected("Sound")
stime = 0
etime = Get total duration

#intesity analysis
To Intensity... 'minpitch' 'tstep' no
intensity = selected("Intensity")
numframes = Get number of frames

#create a textgrid
Create TextGrid...   'stime' 'etime' sounds
textgrid = selected("TextGrid")
Set interval text... 1 1 p

clearinfo

#initially mark each frame as silent or speech
for i from 1 to numframes
      select intensity
      value = Get value in frame... 'i'
      time = Get time from frame number... 'i'
      select textgrid
      Insert boundary... 1 'time'

      interval = Get interval at time... 1 'time'
      if value > nthresh
            label$ = smark$
      else
            label$ = pmark$
      endif
      Set interval text... 1 'interval' 'label$'
endfor

select intensity
Remove
#connect adjacent intervals
call joinadjacent
printline 1

#now eliminate sounds shorter than the threshohld
```

```
call eliminate 'smark$' 'pmark$' 'minsound'
#connect adjacent intervals
call joinadjacent
printline 2

#and finally eliminate pauses shorter than the threshold
call eliminate 'pmark$' 'smark$' 'minpause'
#connect adjacent intervals
call joinadjacent
printline 3

select textgrid


procedure joinadjacent
      select textgrid
      Duplicate tier... 1 2 sounds
      numintervals = Get number of intervals... 1
      for i from 2 to numintervals
            lab1$ = Get label of interval... 1 'i'-1
            lab2$ = Get label of interval... 1 'i'
            if lab1$ = lab2$
                  btime = Get starting point... 1 'i'
                  Remove boundary at time... 2 'btime'
                  interval = Get interval at time... 2 'btime'
                  label$ = Get label of interval... 2 'interval'
                  label$ = left$(label$,1)
                  Set interval text... 2 'interval' 'label$'
            endif
      endfor
      Remove tier... 1
endproc


procedure eliminate mark1$ mark2$ thresh
      select textgrid
      numintervals = Get number of intervals... 1
      for i from 1 to numintervals
            label$ = Get label of interval... 1 'i'
            if label$ = mark1$
                  istart = Get starting point... 1 'i'
                  iend =  Get end point... 1 'i'
                  idur = iend - istart
                  if idur < thresh
                        Set interval text... 1 'i' 'mark2$'
                  endif
            endif
      endfor
endproc
```

2. A/p feature extraction (first stage)

The first stage of a/p feature extraction is performed as a batch process over intervals marked as speech. The implementation comprises two scripts. The first script (batch script – Listing C.2) loops through the textgrid intervals in order to identify which ones are to be analyzed based on the labels. The second script (library script – Listing C.3) contains feature extraction procedures that are called from the batch script. The inputs to this script are (1) interval selection parameters (interval markers in order to selectively analyze desired labels and interval range in order to analyze only a part of a large file) and (2) a/p feature analysis parameters (pitch detection and intensity analysis parameters are gender-specific). The output of the batch script is a Table object with the intervals and vowels in each interval (detected automatically with the method described in appendix B) as rows and various prosodic features as columns. In order for the batch script to be executed, the sound file must be loaded as a LongSound object and the TextGrid must be converted into a Table object. Both the LongSound and Table objects must be selected together in the Praat object window.

Figure C.1 – Feature extraction input parameters window



Listing C.2: Batch script

```
include analyzer3.praat
form Analysis Settings
      optionmenu Audio_channel: 1
            option Left
            option Right
      natural First_interval 1
      integer Last_interval 0 (=the very last interval)
      boolean Analyze_only_marked_intervals: 1
      sentence Interval_markers
      optionmenu Preserve_times_when_extracting_clips: 1
            option yes
            option no
      comment Pitch analysis settings
      natural Pitch_floor_(Hz) 75
      natural Pitch_ceiling_(Hz) 250
      real time_step_(s) 0.005
      comment Intensity Analysis Settings
      natural Intensity_pitch_floor_(Hz) 75
      real Intensity_time_step_(s) 0.005
endform

#assign values to params
channel$ = audio_channel$
fint = first_interval
lint = last_interval
imark = analyze_only_marked_intervals
if imark = 1
      markers$ = interval_markers$
endif
tstamp$ = preserve_times_when_extracting_clips$
pfloor = pitch_floor
pceil = pitch_ceiling
tstep = time_step
ifloor = intensity_pitch_floor
istep = intensity_time_step

#show the info window
clearinfo
printline channel 'channel$'
printline tier 'tier'
printline intervals 'fint' to 'lint'
if imark = 1
      printline interval markers: 'markers$'
endif
printline Preserve times when extracting clips: 'tstamp$'
printline pitch range 'pfloor' - 'pceil' Hz every 'tstep' seconds
printline intensity frame pitch floor 'ifloor' every 'istep' seconds
pause

#start extracting sounds and analyzing
longsound = selected ("LongSound")
labeltable=selected("Table")
select labeltable
Append column... nclip
select all
minus longsound
minus labeltable
nocheck Remove
select labeltable
```

```
Create Table with column names... table 1 clip vowel text start end pmin tpmin
   pmax tpmax pmean pstd imin timin imax timax imean istd jitter shimmer hnr vbr
data = selected("Table")
crow = Get number of rows
select labeltable
intervals = Get number of rows
if lint > intervals
      lint=intervals
elsif lint <= 0
      lint = intervals
endif
intervals = lint-fint+1
printline analyzing 'intervals' intervals

#main loop
nsound=0
for n from 'fint' to 'lint'
      if nsound >= 1
      #pause Do you want to continue?
      endif
      select labeltable
      intervlabel$=Get value... 'n' label
      if imark = 0
            call Newclip
      else
            if index(markers$,intervlabel$)>0
                  call Newclip
            endif
      endif
endfor

printline Analysis of 'intervals' intervals complete. Found 'nsound'clips.
if nsound > 0
      select data
      crow = Get number of rows
      Remove row... 'crow'
      Edit
endif

procedure Newclip
      nsound=nsound+1
      intervalstart = Get value... 'n' start
      intervalend = Get value... 'n' end
      duration = intervalend - intervalstart
      select labeltable
      Set numeric value... 'n' nclip 'nsound'
      select data
      Set numeric value... 'crow' clip 'nsound'
      if duration < 0.25
            Set numeric value... 'crow' vowel -1
      else
            Set numeric value... 'crow' vowel 0
      endif
      Set string value... 'crow' text 'intervlabel$'
      Set numeric value... 'crow' start 'intervalstart'
      Set numeric value... 'crow' end 'intervalend'
      select longsound
      Extract part... 'intervalstart' 'intervalend' 'tstamp$'
      Rename... sound'nsound'
      soundid = selected("Sound")
      printline Analyzing clip 'nsound' ...
```

```
      select data
      vw = Get value... 'crow' vowel
      if vw = 0
            call Analyzer soundid data 'nsound'
      endif
      select soundid
      Remove
      printline clip 'nsound' completed
      select data
      crow = Get number of rows
      Append row
      crow = crow+1
endproc
```

Listing C.3: Library script

```
procedure Analyzer sound table clip
	#clearinfo
	select table
	row = Get number of rows

	#Voice quality of whole clip
	select sound
	noprogress To Pitch (cc)... 'tstep' 'pfloor' 15 no 0.03 0.45 0.01 0.35
	 0.14 'pceil'
	pitchVQ = selected("Pitch")
	select sound
	plus pitchVQ
	noprogress To PointProcess (cc)
	pointprocessVQ = selected("PointProcess")
	call AnalyzeVQ 0 0 sound pitchVQ pointprocessVQ
	#Update table
	select table
	if jitt = undefined
		#do nothing
	else
		jitt=round(jitt*100000)/100000
		Set numeric value... 'row' jitter 'jitt'
	endif
	if shim = undefined
		#do nothing
	else
		shim=round(shim*100000)/100000
		Set numeric value... 'row' shimmer 'shim'
	endif
	if harm2noise = undefined
		#do nothing
	else
		Set numeric value... 'row' hnr 'harm2noise'
	endif
	if vbreaks = undefined
		#do nothing
	else
		vbreaks = round(vbreaks*10000)/100
		Set numeric value... 'row' vbr 'vbreaks'
	endif
	print VQ OK...

	#Pitch of whole clip
	select sound
	noprogress To Pitch... 'tstep' 'pfloor' 'pceil'
	  pitchPI = selected("Pitch")
	call AnalyzePI 0 0 pitchPI
	  #Update table
	select table
	if min = undefined
		#do nothing
	else
		min=round(min*10)/10
		Set numeric value... 'row' pmin 'min'
	endif
	if tmin = undefined
		#do nothing
	else
		tmin=round(tmin*1000)/1000
```

```
      Set numeric value... 'row' tpmin 'tmin'
endif
if max = undefined
      #do nothing
else
      max=round(max*10)/10
      Set numeric value... 'row' pmax 'max'
endif
if tmax = undefined
      #do nothing
else
      tmax=round(tmax*1000)/1000
      Set numeric value... 'row' tpmax 'tmax'
endif
if mean = undefined
      #do nothing
else
      mean=round(mean*10)/10
      Set numeric value... 'row' pmean 'mean'
endif
if std = undefined
      #do nothing
else
      std=round(std*10)/10
      Set numeric value... 'row' pstd 'std'
endif
print PI OK...

#Intensity of whole clip
select sound
To Intensity... 'ifloor' 'istep'
intensityPI = selected("Intensity")
call AnalyzeINT 0 0 intensityPI
#Update table
select table
if min = undefined
      #do nothing
else
      min=round(min*10)/10
      Set numeric value... 'row' imin 'min'
endif
if tmin = undefined
      #do nothing
else
      tmin=round(tmin*1000)/1000
      Set numeric value... 'row' timin 'tmin'
endif
if max = undefined
      #do nothing
else
      max=round(max*10)/10
      Set numeric value... 'row' imax 'max'
endif
if tmax = undefined
      #do nothing
else
      tmax=round(tmax*1000)/1000
      Set numeric value... 'row' timax 'tmax'
endif
if mean = undefined
      #do nothing
```

```
      else
            mean=round(mean*10)/10
            Set numeric value... 'row' imean 'mean'
      endif
      if std = undefined
            #do nothing
      else
            std=round(std*10)/10
            Set numeric value... 'row' istd 'std'
      endif
      print INT OK...
      # Detect Vowels
      call DetectVowels sound table clip
      printline number of vowels: 'nvowels'

      select pitchVQ
      plus pointprocessVQ
      plus pitchPI
      plus intensityPI
      Remove
endproc

procedure AnalyzePI t1 t2 pitch
      select pitch
      min = Get minimum... 't1' 't2' Hertz Parabolic
      tmin = Get time of minimum... 't1' 't2' Hertz Parabolic
      max = Get maximum... 't1' 't2' Hertz Parabolic
      tmax= Get time of maximum... 't1' 't2' Hertz Parabolic
      mean = Get mean... 't1' 't2' Hertz(logarithmic)
      std = Get standard deviation... 't1' 't2' Hertz(logarithmic)
endproc

procedure AnalyzeINT t1 t2 intensity
      select intensity
      min = Get minimum... 't1' 't2' Parabolic
      tmin = Get time of minimum... 't1' 't2' Parabolic
      max = Get maximum... 't1' 't2' Parabolic
      tmax = Get time of maximum... 't1' 't2' Parabolic
      mean = Get mean... 't1' 't2' energy
      std = Get standard deviation... 't1' 't2'
endproc

procedure AnalyzeVQ t1 t2 sound pitch pointprocess
      select sound
      plus pitch
      plus pointprocess
      voicereport$ = Voice report... 't1' 't2' 75 600 1.3 1.6 0.03 0.45
      jitt = extractNumber (voicereport$,"Jitter (local): ")
      shim = extractNumber (voicereport$,"Shimmer (local): ")
      harm2noise = extractNumber(voicereport$,"Mean harmonics-to-noise ratio: ")
      vbreaks = extractNumber (voicereport$,"Degree of voice breaks: ")
endproc

procedure DetectVowels*

endproc
```

* Procedure DetectVowels is the method described in appendix B and is available online as a separate script (see appendix B)

3. A/p feature extraction (second stage)

The output of the first stage is raw data in the form of a table as shown below. The second stage processes this table in order to extract summary information on each speech interval in a more presentable form. This script takes no input arguments. The raw table from the first stage must be selected in the Praat object window.

FIGURE C.2 – Raw feature extraction data

row	start	end	pmin	tpmin	pmax	tpmax
1	106.38	106.91599999999998	157.11657858532396	0.17742329418840924	232.3512504579838	0.42455813116835395
2	0.14262990848749116	0.18553812292982427	157.11657858532396	0.17742329418840924	159.1455856181241	0.18553812292982427
3	0.3095026871199677	0.3437384945809843	221.30155497771813	0.3437384945809843	224.70482084972008	0.3095026871199677
4	0.37789894560806225	0.3786681654348115	224.06966738760215	0.3786681654348115	224.1512595520766	0.37789894560806225

Script functions:

1) Vowel enumeration: The script enumerates the vowels in each speech interval
2) Duration: The script calculates the duration of each speech interval
3) Speech rate: speech rate is estimated as number of vowels/min (see appendix B)
4) Pitch range: the script calculates the pitch range (see section 6.5.3)
5) Vowel-only based a/p measurements (experimental feature). The script calculates the average pitch, intensity, pitch range based on the vowels only (using Equation 7.3)
6) Vowel duration based speech rate estimation (experimental feature): The script calculates the average vowel duration of an interval as an additional estimate of speech rate

The output of the script(Listing C.4) is a new table which contains the above measurements.

Listing C.4: Table process script

```
clearinfo
tablein = selected("Table")
Copy...  newtable
tableout = selected("Table")
Append column... duration
Append column... prange
Append column... speed
Append column... vpmean
Append column... vprange
Append column... vimean
Append column... avd
select tablein
rows = Get number of rows
for i from 1 to 'rows'
      select tablein
      nv = Get value... 'i' vowel
      if nv = 0
            clip = Get value... 'i' clip
            start = Get value... 'i' start
            end = Get value... 'i' end
            pstd = Get value... 'i' pstd
            #pmean = Get value... 'i' pmean
            #imean = Get value... 'i' imean
            duration = end-start
            duration = round(duration*1000)/1000
            prange = 2*pstd
            select tableout
            Set numeric value... 'i' duration 'duration'
            if prange = undefined
                  prange=0
            else
                  prange=round(prange*10)/10
            endif
            Set numeric value... 'i' prange 'prange'
            select tablein
            Extract rows where column (text)... clip "is equal to" 'clip'
            tabletemp1=selected("Table")
            nvowels=Get number of rows
            nvowels=nvowels-1
            if nvowels>0
                  Extract rows where column (number)... vowel "greater than" 0
                  tabletemp = selected("Table")
                  nvowels = Get number of rows
                  vpmean=0
                  vprange=0
                  vimean=0
                  vtdur=0
                  for j from 1 to nvowels
                        vstart=Get value... 'j' start
                        vend=Get value... 'j' end
                        vdur=vend-vstart
                        vdur=round(vdur*1000)/1000
                        vtdur=vtdur+vdur
                        vpm=Get value... 'j' pmean
                        if vpm = undefined
                              #do nothing
                        else
                              vpmean=vpmean+vpm*vdur
                        endif
```

```
                  vpmin=Get value... 'j' pmin
                  vpmax=Get value... 'j' pmax
                  if vpmin = undefined or vpmax = undefined
                        #do nothing
                  else
                        vprange=vprange+(vpmax-vpmin)*vdur
                  endif
                  vim=Get value... 'j' imean
                  if vim = undefined
                        #do nothing
                  else
                        vimean=vimean+vim*vdur
                  endif
            endfor
            vpmean=vpmean/vtdur
            vpmean=round(vpmean*10)/10
            vimean=vimean/vtdur
            vimean=round(vimean*10)/10
            vprange=vprange/vtdur
            vprange=round(vprange*10)/10
            avd=vtdur/nvowels
            avd=round(avd*1000)/1000
            select tabletemp
            plus tabletemp1
            Remove
      else
            select tabletemp1
            Remove
      endif
      speed = 60 * (nvowels / duration)
      speed=round(speed*10)/10
      select tableout
      Set numeric value... 'i' vowel 'nvowels'
      Set numeric value... 'i' speed 'speed'
      if vpmean = undefined
            vpmean=0
      endif
      Set numeric value... 'i' vpmean 'vpmean'
      if vimean = undefined
            vimean=0
      endif
      Set numeric value... 'i' vimean 'vimean'
      if vprange = undefined
            vprange=0
      endif
      Set numeric value... 'i' vprange 'vprange'
      if avd = undefined
            avd=0
      endif
      Set numeric value... 'i' avd 'avd'
      printline 'i' 'duration' 'nvowels' 'pmean' 'imean' 'prange' 'speed'
       'vpmean' 'vprange' 'vimean'  'avd'
elsif nv=-1
      select tableout
      Set numeric value... 'i' vowel 0
      start = Get value... 'i' start
      end = Get value... 'i' end
      duration = end-start
      duration = round(duration*1000)/1000
      select tableout
      Set numeric value... 'i' duration 'duration'
```

```
      else
            select tableout
            Set numeric value... 'i' vowel -1
      endif
endfor

select tableout
Extract rows where column (number)... vowel "greater than or equal to" 0
finaltable = selected("Table")
select tableout
Remove

#replace NaN with 0s
select finaltable
n=Get number of rows
for i from 1 to n
      nv = Get value... 'i' vowel
      if nv = 0
            for j from 6 to 28
                  col$ = Get column label... 'j'
                  value = Get value... 'i' 'col$'
                  if value = undefined
                        Set numeric value... 'i' 'col$' 0
                  endif
            endfor
      endif
endfor
```

4. TAMA analysis

TAMA analysis (see chapter 7) is performed in MATLAB software. The output tables of the feature extraction (for each of the two speakers) are imported into MATLAB as tab-delimited table files. In MATLAB, these are represented as matrices of numbers, in which rows correspond to speech intervals and columns correspond to prosodic features. The first step in TAMA analysis is to acquire TAMA feature vectors, which are essentially the univariate time series used in the statistical analysis. The *tamaframe* function (Listing C.5) takes an imported data table as input and outputs a TAMA vector of the desired a/p feature (column).

The inputs to this function are (1) the input matrix (Praat imported table), (2) the column number (desired a/p feature), (3) TAMA frame length and time step (see section 7.3.1), and (4) duration limits (minimum and maximum) if it is desirable to ignore intervals above/below a certain duration. The function outputs a new 4-column matrix. The first two columns are the start and end times of the frames, which can be used as indices in TAMA plots. The third column is the a/p feature vector, and the fourth column contains the relative duration for each frame (see section 7.3.1).

Setting the frame length and time step equal to or greater than the duration of the dialogue yields the grand mean of the desired feature. Dividing a TAMA vector by this value yields the normalized values (with mean equal to 1) of that feature.

The feature vector is then extracted from the output matrix and used as a component series in the statistical analysis (bi-variate time series), as described in section 7.4.

Listing C.5: Function Tamaframe

```
%This function creates a TAMA frame vector based on the input parameters

function [TAMA] =
tamaframe(matrix,columnnumber,framelength,timestep,mindur,maxdur)

%calculate number of frames
n = length(matrix);
lastend = matrix(n,2);
numberofframes = fix(lastend/timestep)+1;

% main loop
for i=1:numberofframes
    %calculate frame boundaries
    framestart=(i-1)*timestep;
    frameend = framestart + framelength;
    % set end of frame flag
    endofframe=0;
    k = 1;  %matrix row index
    wsum = 0;   %weighted sum initialization
    sdur = 0;   %duration sum initialization
    while endofframe==0
        %find clips within frame
        clipstart = matrix(k,1);
        if clipstart > frameend
            endofframe = 1;
        end
        clipend = matrix(k,2);
        %clip intervals at frame boundaries
        if clipstart < framestart
            clipstart = framestart;
        end
        if clipend > frameend
            clipend = frameend;
        end
        duration = clipend-clipstart
        %is interval in the frame?
        if duration > 0
            %duration limits check
            if (matrix(k,19)>=mindur)&&(matrix(k,19)<=maxdur)
                %check for NaN
                if (isfinite(matrix(k,columnnumber))==1)
                  %add to weigthed sum
                  wsum = wsum + matrix(k,columnnumber)*duration;
                  %add to duration sum
                  sdur = sdur + duration;
                end
            end
        end
        k = k + 1;
        if k > n
            endofframe = 1;
        end
    end
    TAMA(i,1) = framestart;
    TAMA(i,2) = frameend;
    TAMA(i,3) = wsum/sdur;
    TAMA(i,4) = sdur;
end
```

5. Acquisition of combined chronograph

The TextGrid objects acquired during the segmentation phase (manually corrected) are essentially the individual chronographs (see Figure 8.1) of the two speakers. In order to acquire the combined chronograph (Figure 8.2) of the dialogue, the two individual chronographs have to be superimposed on each other. This is performed by the following Praat script (Listing C.6). The input to this script are two Table objects, which can be easily obtained in recent Praat versions by selecting a TextGrid object and using the appropriate command from the dynamic menu. Both these tables have to be selected in the object list before the script is executed. The output of the script is a new textgrid with four different labels (speaker 1, speaker 2, pause, overlap). This textgrid is also converted to a table using the built-in command.

Listing C.6 - Chronograph script

```
table1=selected("Table",1)
table2=selected("Table",2)

#copy boundaries from both tabels into new textgrid
select table1
n1=Get number of rows
maxtime1=Get value... 'n1' end
select table2
n2=Get number of rows
maxtime2=Get value... 'n2' end
maxtime=max(maxtime1,maxtime2)
Create TextGrid... 0 'maxtime' interval
mtextgrid=selected("TextGrid")
call Table2textgrid table1 mtextgrid
call Table2textgrid table2 mtextgrid

clearinfo
printline copied boundaries

#compare labels and set new labels for combined chronograph
select mtextgrid
n = Get number of intervals... 1
for i from 1 to n
      start=Get start point... 1 i
      end=Get end point... 1 i
      dur=end-start
      mid=start+dur/2
      #Find the intervals in both label tables
      call Fintv table1 'mid'
      print 'mid','interval',
      lab1$ = Get value... 'interval' label
      call Fintv table2 'mid'
      print 'interval'
      printline
      lab2$ = Get value... 'interval' label
      if lab1$="p"
            lab1$=""
      else
            lab1$="t1"
      endif
      if lab2$="p"
            lab2$=""
      else
            lab2$="t2"
      endif
      lab$=lab1$+lab2$
      if lab$=""
            lab$="p"
      endif
      if length(lab$)>2
            lab$="o"
      endif
      select mtextgrid
      Set interval text... 1 'i'  'lab$'
endfor

#connect neibhouring labels that are equal
select mtextgrid
Copy...  turndist
```

```
turndist=selected("TextGrid")
endtime = Get end time
i=1
repeat
      i=i+1
      endinterval = Get end point... 1 i
      lab1$ = Get label of interval... 1 i
      lab2$ = Get label of interval... 1 i-1
      if lab1$=lab2$
            Remove left boundary... 1 i

            Set interval text... 1 i-1  'lab1$'
            i=i-1
      endif
until endinterval=endtime

print ok
select mtextgrid
Remove
select turndist
Rename... timeline

procedure Table2textgrid table textgrid
      select table
      .n=Get number of rows
      for i from 1 to '.n'
            select table
            t=Get value... 'i' start
            select textgrid
            nocheck Insert boundary... 1 't'
      endfor
endproc

procedure Fintv table time
      interval=0
      .i = 1
      select table
      while interval=0
            t1=Get value... '.i' start
            t2=Get value... '.i' end
            if time>t1&&time<t2
                  interval='.i'
            endif
            .i=.i+1
      endwhile
endproc
```

6. Pause and overlap annotation algorithm

The algorithm for switch pause and overlap annotation described in section 8.2.1 is implemented as a MATLAB script. The chronograph tables (individual speaker and combined) are imported as matrices of numbers, in which rows correspond to the table rows (or TextGrid intervals). The matrices have three columns. The first two contain the start and end time of each interval, and the the third column corresponds to the interval label. The following table shows the label-to-number conversion.

Table C.1 – Label-to-number conversion

Chronograph type	Individual	Combined
Number	Label	Label
1	-	Speaker 1
2	-	Speaker 2
3	-	Overlap
4	Speech	-
5	-	-
6	other	other
7	Silence	Silence

The individual chronograph tables are used to calculate the speakers' turn shares (see section 8.3.1), while the combined chronograph table is used to acquire the turn share distribution (section 8.3.1), as well as to implement the switch pause algorithm described in section 8.2.1, using the following MATLAB script (Listing C.7). The input to the script is the combined chronograph (labeltableTn), and the output comprises three two-column matrices. The first column in these matrices contains the time instant at which a turn-switch occurs, and the second column contains the duration of the preceding pause. Two of the three matrices correspond to speaker 1 and speaker 2, while the third matrix contains the ambiguous cases (simultaneous starts after pauses) which are very rare.

Listing C.7: Switch pause detection script

```
%create switch pause tables from label table T


n = length(labeltableTn);
spauseL = [];
spauseR = [];
spauseamb = [];
for i = 2:(n-1)
    switch labeltableTn(i,3)
        case 7
            if labeltableTn(i-1,3)~=labeltableTn(i+1,3)
                switch labeltableTn(i+1,3)
                    case 1
                        k = length(spauseL);
                        k = k + 1;
                        spauseL(k,1) = labeltableTn(i,1);
                        spauseL(k,2) = labeltableTn(i,2)-labeltableTn(i,1);
                    case 2
                        k = length(spauseR);
                        k = k +1;
                        spauseR(k,1) = labeltableTn(i,1);
                        spauseR(k,2) = labeltableTn(i,2)-labeltableTn(i,1);
                    otherwise
                        k = length(spauseamb);
                        k = k +1;
                        spauseamb(k,1) = labeltableTn(i,1);
                        spauseamb(k,2) = labeltableTn(i,2)-labeltableTn(i,1);
                end
            end
        case 3
            if labeltableTn(i-1,3)~=labeltableTn(i+1,3)
                 switch labeltableTn(i+1,3)
                    case 1
                        k = length(spauseL);
                        k = k + 1;
                        spauseL(k,1) = labeltableTn(i,1);
                        spauseL(k,2) = 0;
                    case 2
                        k = length(spauseR);
                        k = k +1;
                        spauseR(k,1) = labeltableTn(i,1);
                        spauseR(k,2) = 0;
                    otherwise
                        k = length(spauseamb);
                        k = k +1;
                        spauseamb(k,1) = labeltableTn(i,1);
                        spauseamb(k,2) = 0;
                 end
            end
    end
end
```

List of Publications

Kousidis, S., D. Dorran, C. McDonnell and E. Coyle (2009b). Towards Flexible Representations for Analysis of Accomodation of Temporal Features in Spontaneous Dialogue Speech. InterSpeech 2009. Brighton, United Kingdom.

Kousidis, S., D. Dorran, C. Mcdonnell and E. Coyle (2009a). Time Series Analysis of Acoustic Feature Convergence in Human Dialogues. SPECOM 2009. St Petersburg, Russian Federation.

Kousidis, S. and D. Dorran (2009). Monitoring Convergence of Temporal Features in Spontaneous Dialogue Speech. 1st Young Researchers Workshop on Speech Technology. UCD, Dublin, Ireland.

Kousidis, S., D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, c. McDonnell and E. Coyle (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues Interspeech 2008. Brisbane, Australia.

Cullen, C., Vaughan, B., Kousidis, S. (2008a). Emotional Speech Corpus Construction, Annotation and Distribution. The 6th edition of the Language Resources and Evaluation Conference. Marrakech (Morocco).

Cullen, C., Vaughan, B., Kousidis, S. (2008b). LinguaTag: an emotional speech analysis application. The 12th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2008. Orlando, Florida, USA.

Cullen, C., B. Vaughan, S. Kousidis, Y. Wang, C. McDonnell and D. Campbell (2006). Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain.

References

Adda-Decker, M., C. Barras, G. Adda, P. Paroubek, P. B. d. Mareüil and B. Habert (2008). Annotation and analysis of overlapping speech in political interviews The 6th edition of the Language Resources and Evaluation Conference (LREC 2008). Marrakech (Morocco). .

Allen, J. F., D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu and A. Stent (2001). "Toward Conversational Human-Computer Interaction " AI Magazine **22**(4).

Allen, J. F., C. I. Guinn and E. Horvtz (1999). "Mixed-initiative interaction." Intelligent Systems and their Applications, IEEE **14**(5): 14-23.

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta and P. Paggio (2007). "The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena." Language Resources and Evaluation **41**(3): 273-287.

Altman, I., A. Vinsel and B. Brown (1981). Dialectic Conceptions in Social Psychology: An Application to Social Penetration and Privacy Regulation. Advances in Experimental Social Psychology. L. Berkowitz. New York, USA, Academic Press.

Andersen, P. A. (1999). Nonverbal communication: Forms and functions. Mountain View, CA, USA, Mayfield Publishing.

Ang, J., Y. Liu and E. Shriberg (2005). Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), Philadelphia, PA, USA.

Argyle, M. and J. Dean (1965). "Eye-contact, distance and affiliation." Sociometry **28**(3): 289-304.

Aronson, W. A. (2007). Social Psychology. New Jersey, USA, Pearson Education.

Askenfelt, A., J. Gauffin and J. Sundberg (1980). "A Comparison of Contact Microphone and Electroglottograph for the Measurement of Vocal Fundamental Frequency." Journal of Speech and Hearing Research **23**: 258-273.

Aubergé, V. (2002). A Gestalt morphology of prosody directed by functions : the example of a step by step model developed at ICP. 1st International Conference on Speech Prosody, Aix-en-Provence, France.

Austermann, A., N. Esau, L. Kleinjohann and B. Kleinjohann (2005). Fuzzy emotion recognition in

natural speech dialogue. IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005.

Banse, R. and K. R. Schrerer (1996). "Acoustic profiles in vocal emotion expression." Journal of personality and social psychology **70**(3): 614-636.

Barbosa, P. A. (2009). Measuring speech rhythm variation in a model-based framework. Interspeech 2009. Brighton, United Kingdom.

Batliner, A., K. Fischer, R. Huber, J. Spilker and E. Noth (2000). Desperately Seeking Emotions or: Actors, Wizards, and Human Beings. ISCA workshop on Speech and Emotion, Northern Ireland.

Bavelas, J. B., A. Black, C. R. Lemery and J. Mullett (1986). ""I show how you feel": Motor mimicry as a communicative act." Journal of Personality and Social Psychology **50**(2): 322-329.

Beattie, G. W. (1982). "Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted." Semiotica **39**(1): 93-114.

Bell, L., J. Boye, J. Gustafson and M. Wirén (2000). Modality Convergence in a Multimodal Dialogue System. Götalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue, Göteborg University, Sweden.

Bell, L., J. Gustafson and M. Heldner (2003). Prosodic adaptation in human-computer interaction. ICPhS, Barcelona.

Benus, S. (2009). Are we 'in sync': Turn-taking in collaborative dialogues. Interspeech 2009. Brighton, UK.

Bernieri, F. and R. Rosenthal (1991). Coordinated movement in human interaction. Fundamentals of nonverbal behavior. F. Rime. New York, Cambridge University Press**:** 401-432.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. . IFA.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer [computer program].

Bomsdorf, B. and G. Szwillus (1999). From Task to Dialogue Modelling Based on a Tool-Supported Framework. CHI'99 Workshop "Tool Support for Task-Based User Interface Design". Pittsburgh/PA, USA.

Bosch, L. t., N. Oostdijk and L. Boves (2005). "On temporal aspects of turn taking in conversational

dialogues." Speech Communication **50**(1-2): 80-86.

Bosch, L. t., N. Oostdijk and J. P. d. Ruiter (2004a). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. 7th International Conference TSD 2004, Brno, Czech Republic.

Bosch, L. T., N. Oostdijk and J. P. D. Ruiter (2004b). Turn-taking in social talk dialogues: Temporal, formal and functional aspects. SPECOM, St Petersburg.

Botinis, A., B. Grantstrom and B. Mobius (2001) "Developments and paradigms in intonation research." Speech Communication **33**, 263-296.

Boves, L. and E. d. Os (1999). Applications of Speech Technology: Designing for Usability. IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, Co,USA.

Brennan, S. E. (1996). Lexical Entrainment in Spontaneous Dialogue. International Symposium on Spoken Dialogue, ISSD-96, Philadelphia, PA, USA, Acoustical Society of Japan.

Brennan, S. E. and H. H. Clark (1996). "Conceptual pacts and lexical choice in conversation." Journal of Experimental Psychology: Learning, Memory and Cognition **22**(6): 1482-1493.

Bruce, G., J. Frid, B. Granstrom, K. Gustafson, M. Horne and D. House (1996). "Prosodic Segmentation and structuring of dialogue." TMH-QPSR **37**(3): 1-6.

Buder, E. H. and A. Eriksson (1997). Prosodic cycles and interpersonal synchrony in American English and Swedish. EUROSPEECH-1997, Rhodes, Greece.

Buder, E. H. and A. Eriksson (1999). Time-series analysis of conversational prosody for the identification of rhythmic units. The XIVth International Congress of Phonetic Sciences, San Francisco.

Burgoon, J. K. (1978). "A communication model of personal space violations: explication and an initial test. ." Human Communication Research. **4**: 129-142.

Burgoon, J. K., L. A. Stern and L. Dillman (1995). Interpersonal Adaptation: Dyadic Interaction Patterns, Cambridge university Press.

Campbell, N. (2000). Databases of emotional speech. ISCA Workshop on Speech and Emotion, Northern Ireland.

Campbell, N. (2006). "Conversational speech synthesis and the need for some laughter." IEEE Transactions on Audio, Speech, and Language Processing **14**(4): 1171-1178.

Campbell, N. (2009). An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal

Conversation Data. Interspeech 2009. Brighton ,UK.

Cappella, J. N. and J. O. Green (1982). "A discrepancy-arousal explanation of mutual influence in expressive behavior for adult and infant-infant interaction." Communication Monographs(49): 89-114.

Carlson, R., J. Edlund, M. Heldner, A. Hjalmarsson, D. House and G. Skantze (2006). Towards human-like behaviour in spoken dialog systems. Swedish Language Technology Conference (SLTC). Gothenburg, Sweden.

Cerrato, L. (2002). Some characteristics of feedback expressions in Swedish. Fonetik 2002, Stockholm, Sweden.

Chang, S., L. Shastri and S. Greenberg (2000). Automatic phonetic transcription of spontaneous speech (American English). International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China.

Chatfield, C. (1996). The Analyis of Time Series - An Introduciton, Chapman & Hall/CRC.

Chu, M., H. Peng and E. Chang (2001). A concatenative Mandarin TTS system without prosody model and prosody modification. 4th ISCA workshop on speech synthesis, Scotland.

Clark, H. H. and S. E. Brennan (1991). Grounding in communication. Perspectives on socially shared cognition. L. B. Resnick, J. M. Levine and S. D. Teasley. Washington, DC, US, American Psychological Association**:** 127-149.

Clark, H. H. and E. F. Schaefer (1989). "Contributing to Discourse." Cognitive Science(13): 259-294.

Clark, H. H. and D. Wilkes-Gibbs (1986). "Referring as a collaborative process." Cognition **22**(1): 1-39.

Condon, W. and W. Ogston (1971). Speech and body motion synchrony of the speaker-hearer. The Perception of Language. D. H. a. J. Jenkins, 1971. **:** 150-184.

Condon, W. S. and W. D. Ogston (1966). "Sound film analysis of normal and pathological behavior patterns." Journal of Nervous and Mental Disease **143**(4).

Condon, W. S. and W. D. Ogston (1967). "A segmentation of behavior." Journal of Psychiatric Research **5**(3): 221-235.

Cornelius, R. R. (2000). Theoretical Approaches to Emotion. ISCA Workshop on Speech and Emotion, Belfast, northern Ireland.

Coulston, R., S. Oviatt and C. Darves (2002). Amplitude convergence in children²s conversational speech with animated personas. International conference on spoken language processing (ICSLP-2002), Denver, Colorado, USA.

Cowie, R. and R. R. Cornelius (2003). "Describing the emotional states that are expressed in speech." Speech Communication Special Issue on Speech and Emotion **40**(1-2): 5-32.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor (2001). "Emotion recognition in human-computer interaction." IEEE Signal Processing Magazine **18**(1): 32-80.

Cullen, C., B. Vaughan, S. Kousidis, Y. Wang, C. McDonnell and D. Campbell (2006). Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction International Conference on Multidisciplinary Information Sciences and Technologies Extremadura, Merida.

Cullen, C., Vaughan, B., Kousidis, S. (2008a). Emotional Speech Corpus Construction, Annotation and Distribution. The 6th edition of the Language Resources and Evaluation Conference. Marrakech (Morocco).

Cullen, C., Vaughan, B., Kousidis, S. (2008b). LinguaTag: an emotional speech analysis application. The 12th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2008. Orlando, Florida, USA.

Cummins, F. and R. F. Port (1998). "Rhythmic constraints on stress timing in English." Journal of Phonetics **26**(2): 145-171.

Cutler, A., D. Dahan and W. v. Donsellar (1997). "Prosody in the Comprehension of Spoken Language: A Literature Review." Language and Speech(40(2)): 141-201.

d'Alessandro, C. and P. Mertens (1995). "Automatic pitch contour stylization using a model of tonal perception." Computer Speech & Language **9**(3): 257-288(232).

Darves, C. and S. Oviatt (2002). Adaptation of users spoken dialogue patterns in a conversational interface. International Conference on Speech and Language Processinf (ICSLP-2002), Denver, Colorado.

deJong, N. H. and T. Wempe (2007). "Automatic measurement of speech rate in spoken Dutch." ACLC Working Papers **2**(2): 1-50.

Delmonte, R. (2005). Modeling conversational styles in Italian by means of overlaps. Disfluency in

Spontaneous Speech Workshop (DiSS-2005), Aix-en-Provence.

Dutoit, T. (1997). An Introduction to Text-to-Speech Synthesis. Dordrecht, Kluwer Academic Publishers.

Dybkjær, H. and L. Dybkjær (2004). "Modeling Complex Spoken Dialog." Computer **37**(8): 32-40.

Dybkjær, L., N. O. Bernsen and W. Minker (2004). "Evaluation and usability of multimodal spoken language dialogue systems." Speech Communication **43**(1-2): 33-54.

Edlund, J., J. Gustafson, M. Heldnera and A. Hjalmarssona (2008). "Towards human-like spoken dialogue systems." Speech communication **50**(8-9): 630-645.

Edlund, J., M. Heldner and J. Gustafson (2005). Utterance segmentation and turn- taking in spoken dialogue systems. Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. B. Fisseni, H.-C. Schmitz, B. Schröder and P. Wagner. Frankfurt am Main, Germany, Peter Lang. **:** 576-587.

Edlund, J., M. Heldner and J. Gustafson (2006). Two faces of spoken dialogue systems. Interspeech 2006. Pittsburgh, PA, USA.

Edlund, J., M. Heldner and J. Hirschberg (2009). Pause and gap length in face-to-face interaction. InterSpeech 2009. Brighton, UK.

Fais, L. (1996). Lexical accommodation in machine-mediated interactions. Proceedings of the 16th conference on Computational linguistics - Volume 1. Copenhagen, Denmark, Association for Computational Linguistics.

Fernandez, R., T. Lucht, K. Rodriguez and D. Schlangen (2006). INTERACTION IN TASK-ORIENTED HUMAN-HUMAN DIALOGUE: THE EFFECTS OF DIFFERENT TURN-TAKING POLICIES. IEEE Spoken Language Technology Workshop, 2006, Palm Beach.

Fernandez, R. and R. Picard (2000). Modelling drivers' speech under stress. ISCA Workshop on Speech and Emotion, Northern Ireland.

Fujisaki, H. (1992). The Role of Quantitative Modelling in the Study of Intonation. International Symposium on Japanese Prosody.

Furui, S., M. Nakamura, T. Ichiba and K. Iwano (2005). "Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese." Speech Communication **47**(1-2): 208-219.

Galanis, D., V. Darsinos and G. Kokkinakis (1996). Investigating emotional speech parameters for speech synthesis. Third IEEE International Conference on Electronics, Circuits, and

Systems (ICECS '96), Rodos, Greece.

Gerrards-Hesse, A., K. Spies and F. W. Hesse (1994). "Experimental inductions of emotional states and their effectiveness: A review." British Journal of Psychology **85**: 55-78.

Giles, H., N. Coupland and J. Coupland (1992). Accommodation theory: Communication, context and consequence Contexts of accommodation: developments in applied sociolinguistics. Howard Giles, N. Coupland and J. Coupland, Cambridge University Press**:** 1-68.

Giles, H., A. Mulac, J. J. Bradac and P. Johnson (1987). Speech Accomodation Theory: The First Decade and Beyond. Communication Yearbook 10. M. L. McLaughlin. Newbury Park, SAGE**:** 13-48.

Glass, J. R. (1999). Challenges for spoken dialogue systems. IEEE AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING WORKSHOP. Keystone, Colorado, U.S.A.

Gobl, C., E. Bennett and A. N. Chasaide (2002). Expressive Synthesis: How Crucial is Voice Quality? IEEE Workshop on Speech Synthesis, Santa Monica, CA (USA).

Gouldner, A. W. (1960). "The Norm of Reciprocity: A Preliminary Statement." American Sociological Review **25**(2): 161-178.

Gratier, M. (2003). "Expressive timing and interactional synchrony between mothers and infants: cultural similarities, cultural differences, and the immigration experience." Cognitive Development **18**(4): 533-554.

Gross, J. J. and R. W. Levenson (1995). "Emotion elicitation using films." Cognition & Emotion **9**(1): 87-108.

Gussenhoven, C. (2005). The Phonology of Tone and Intonation. Cambridge, Cambridge University Press.

Hakulinen, J. and M. Turunen (1999). Prosodic Features for Speech User Interfaces. ACHCI'99, Report B.

Hardy, H., A. Biermann, R. B. Inouye, A. Mckenzie, T. Strzalkowski, C. Ursu, N. Webb and M. Wu (2004). Data driven strategies for an automated dialogue system. 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona.

Hastie, H. W., M. Poesio and S. Isard (2002). "Automatically predicting dialogue structure using prosodic features." Speech Communication(36): 63-79.

Heylen, D. (2009). Understanding Speaker-Listener Interactions. Interspeech 2009. Brighton, UK.

Hirose, K., M. Sakata and H. Kawanami (1996). Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features. Fourth International Conference on Spoken Language (ICSLP 96).

Holzapfel, H., C. Fuegen, M. Denecke and A. Waibel (2002). Integrating emotional cues into a framework for dialogue management. Fourth IEEE International Conference on Multimodal Interfaces.

Homans, G. C. (1958). "Social Behavior as Exchange." The American Journal of Sociology **63**(6): 597-606.

Jaffe, J., B. Beebe, S. Feldstein, C. L. Crown, M. D. Jasnow, P. Rochat and D. N. Stern (2001). "Rhythms of Dialogue in Infancy: Coordinated Timing in Development." Monographs of the Society for Research in Child Development **66**(2): i-149.

Jaffe, J. and S. Feldstein (1970). Rhythms of Dialogue. New York, Academic Press.

Johnstone, T. (1996). Emotional Speech Elicited Using Computer Games. 4th International Conference on Speech and Language Processing, Philadelphia, PA, USA.

Johnstone, T., C. M. v. Reekum, K. Hird, K. Kirsner and K. R. Scherer (2005). "Affective speech elicited with a computer game." Emotion: 513-518.

Johnstone, T. and K. R. Scherer (1999). The Effects of Emotions on Voice Quality. XIV Int. Congress of Phonetic Sciences, San Francisco.

Jokinen, K. (2000). Learning Dialogue Systems. LREC workshop From Spoken Dialogue to Full Natural Interactive Dialogue, Athens, Greece.

Jokinen, K. (2003). Natural Interaction in Spoken Dialogue Systems. Workshop Ontologies and Multilinguality in User Interfaces.HCI International Crete Greece.

Jourard, S. M. and M. J. Landsman (1960). "Cognition, cathexis, and the dyadic effect in men's self-disclosing behavior." Merrill-Palmer Quarterly **6**: 178-185.

Kakita, K. (1996). Inter-speaker interaction of F0 in dialogs. Fourth International Conference on Spoken Language (ICSLP 96), Wynham Franklin Plaza hotel, Philadelphia, PA, USA.

Kaplan, M. M. and K. J. Kaplan (1984). "A bidimensional view of distancing: Reciprocity versus compensation, intimacy versus social control." Journal of Nonverbal Behavior **8**(4): 315-326.

Katz, R. A. (2002). Mastering Audio: The Art and the Science, Focal Press.

Kehrein, R. (2002). The prosody of authentic emotions. Speech Prosody (SP-2002), Aix-en-Provence, France.

Kochanski, G. (2006). Prosody Beyond Fundamental Frequency. Methods in Empirical Prosody Research S. Sudhoff, D. Lenertova, R. Meyeret al. Berlin, Walter de Gruyter.

Kohler, K. J. (1991). "A model of German intonation." Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)(25): 295-360.

Kohler, K. J. (2004). Prosody Revisited. FUNCTION, TIME , and the LISTENER in Intonational Phonology. Speech Prosody 2004. Nara, Japan.

Kolb, B. and I. Whishaw (2003). Fundamentals of human neuropsychology. New York, NY, USA, Worth Publishers.

Kousidis, S. and D. Dorran (2009). Monitoring Convergence of Temporal Features in Spontaneous Dialogue Speech. 1st Young Researchers Workshop on Speech Technology. UCD, Dublin, Ireland.

Kousidis, S., D. Dorran, C. Mcdonnell and E. Coyle (2009a). Time Series Analysis of Acoustic Feature Convergence in Human Dialogues. SPECOM 2009. St Petersburg, Russian Federation.

Kousidis, S., D. Dorran, C. McDonnell and E. Coyle (2009b). Towards Flexible Representations for Analysis of Accomodation of Temporal Features in Spontaneous Dialogue Speech. InterSpeech 2009. Brighton, United Kingdom.

Kousidis, S., D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, c. McDonnell and E. Coyle (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues Interspeech 2008. Brisbane, Australia.

Kurematsu, A., Y. Akegami, S. Burger, S. Jekat, B. Lause, V. L. Maclaren, D. Oppermann and T. Schultz (2000). VERBMOBIL Dialogues: Multifaced Analysis. International Conference on Speech and Language Processing (ICSLP 2000), Beijing, China.

Ladd, R. D. (1983). "Phonological Features of Intonational Peaks " Language **59**(4): 721-759.

Ladd, R. D. (1996). Intonational Phonology. Cambridge, Cambridge University Press.

Larsson, S. (2005). Dialogue Systems: Simulations or Interfaces? The ninth workshop on the semantics and pragmatics of dialogue (Dialor'05), Nancy, France.

Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge, Cambridge University

Press.

Lee, C. M. and S. S. Narayanan (2005). "Toward detecting emotions in spoken dialogs." IEEE Transactions on Speech and Audio Processing **13**(2): 293-303.

Lehiste, I. (1970). Suprasegmentals. Cambridge, MA, MIT Press.

Lennes, M. and H. Anttila (2002). Prosodic features associated with the distribution of turns in Finnish informal dialogues. The Phonetics Symposium 2002. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing.

Levelt, W. J. M. (1983). "Monitoring and self-repair in speech." Cognition **14**: 41-104.

Litman, D. J. and S. Pan (2002). "Designing and Evaluating an Adaptive Spoken Dialogue System." User Modeling and User-Adapted Interaction (12): 111-137.

Lustgarten, P. C. and B. H. Juang (2003). Naturalness in speech communications. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSSPR-2003). Tokyo, Japan.

Macchi, M. (1998). Issues in Text-to-Speech synthesis. IEEE International Joint Symposia on Intelligence and Systems, Rockville, Maryland.

Maekawa, K., H. Koiso, S. Furui and H. Isahara. (2000). Spontaneous speech corpus of Japanese. 2nd LREC, Athens.

Matessa, M. P. (2001). Interactive Models of Collaborative Communication Twenty-third Annual Conference of the Cognitive Science Society, Hillsdale, NJ, Lawrence Erlbaum Associates.

McRoberts, G. W. and C. T. Best (1997). "Accommodation in mean f0 during mother–infant and father–infant vocal interactions: a longitudinal case study." Journal of Child Language **24**(3): 719-736.

McTear, M. F. (2004). Spoken Dialogue Technology: Toward the Conversational User Interface. London, Springer-Verlag.

Minker, W., A. Albalate, D. Buhler, A. Pittermann, J. Pittermann, P.-M. Strauss and D. Zaykovskiy (2006). Recent Trends in Spoken Language Dialogue Systems. 4th International Conference on Information & Communications Technology (ICICT '06), Cairo, Egypt.

Moller, S., P. Smeele, H. Boland and J. Krebber (2007). "Evaluating spoken dialogue systems according to de-facto standards: A case study." Computer Speech and Language **21**: 26-53.

Murray, I. R. and J. L. Arnott (1993). "Toward the simulation of emotion in synthetic speech: A

review of the literature on human vocal emotion." Journal of the Acoustical Society of America. **93**(2): 1097-1108.

Mushin, I., L. Stirling, J. Fletcher and R. Wales (2003). "Discourse Structure, Grounding, and Prosody in Task-Oriented Dialogue." Discourse Processes **35**(1): 1 - 31.

Nagaoka, C., M. Komori and S. Yoshikawa (2005). Synchrony tendency: interactional synchrony and congruence of nonverbal behavior in social interaction. International Conference on Active Media Technology.

Narayanan, S. and A. Alwan (2004). Text to Speech Synthesis: New Paradigms and Advances. New Jersey, USA, Prentice Hall.

Nishimura, R., N. Kitaoka and S. Nakagawa (2008). Analysis of Relationship Between Impression of Human-to-Human Conversations and Prosodic Change and Its Modeling. Interspeech. Brisbane, Australia.

Oviatt, S., C. Darves and R. Coulston (2004). "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas." ACM Trans. Comput.-Hum. Interact. **11**(3): 300-328.

Oviatt, S. and S. Seneff (2004). "Introduction to Mobile and Adaptive Conversational Interfaces." ACM Transactions on Computer-Human Interaction **11**(3): 237-240.

Pardo, J. S. (2006). "On phonetic convergence during conversational interaction." Journal of the Acoustic Society of America **4**(119): 2382-2393.

Patterson, M. L. (1976). "An arousal model of interpersonal intimacy." Psychological Review **83**(3): 235-245.

Patterson, M. L. (1982). "A sequential functional model of nonverbal exchange." Psychological. Review **89**: 231-249.

Pellegrino, F., J. Farinas and J. L. Rouas (2004). Automatic estimation of speaking rate in multilingual spontaneous speech. Speech Prosody 2004 (SP-2004), Nara, Japan.

Perez-Quinones, M. and J. Sibert (1996). A collaborative model of feedback in human-computer interaction. Proceedings of the SIGCHI conference on Human factors in computing systems: common ground. Vancouver, British Columbia, Canada, ACM.

Persia, L. D., M. Yanagida, H. L. RufinerA and D. Milonea (2007). "Objective quality evaluation in blind source separation for speech recognition in a real room." Signal Processing(87): 1951–

1965.

Pfau, T. and G. Ruske (1998). Estimating the speaking rate by vowel detection. IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, USA.

Picard, R. W., E. Vyzas and J. Healey (2001). "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(10): 1175-1191.

Pickering, M. J. and S. Garrod (2004). "Toward a mechanistic psychology of dialogue." Behavioral and Brain Sciences **27**(2): 169-190.

Pickett, J. M. (1999). The Acoustics of Speech Communication. Toronto, Allyn & Bacon.

Pieraccini, R. and J. Huerta (2005). Where do we go from here? research and commercial spoken dialog systems. SIGdial6-2005.

Pieraccini, R., D. Suendermann, K. Dayanidhi and J. Liscombe (2009). Are We There Yet? Research in Commercial Spoken Dialog Systems. Text, Speech and Dialogue. Berlin, Springer Berlin / Heidelberg**:** 3-13.

Pierrehumbert, J. B. (1980). The Phonology and Phonetics of English Intonation. Dept. of Linguistics and Philosophy. Massachusetts, M.I.T. . **PhD**.

Press, W. H., S. A. Teukolsky and W. T. Vetterling (1992). Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press.

Rabiner, L. R. and R. W. Schafer (1978). Digital Processing of Speech Signals, Prentice Hall.

Rangarajan, V., S. Bangalore and S. Narayanan (2007). Exploiting prosodic features for dialog act tagging in a discriminative modeling framework. INTERSPEECH-2007.

Raux, A. (2008). Flexible Turn-Taking for Spoken Dialog Systems. School of Computer Science. Pittsburgh, PA, USA, Carnegie Mellon University. **PhD**.

Raux, A. and M. Eskenazi (2008). Optimizing end- pointing thresholds using dialogue features in a spoken dialogue system. SIGdial 2008, Columbus, OH, US.

Reitter, D., Johanna D. Moore and F. Keller (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. 28th Annual Conference of the Cognitive Science Society (CogSci), Vancouver, Canada.

Richardson, D. C., R. Dale and K. Shockley (2008). Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. Embodied Communication. G.

Knoblich, Oxford University Press.

Russell, J. A. (1997). How shall an emotion be called? Circumplex Models of Personality and Emotion. R. Plutchik and H. Conte. Washington, APA**:** 205–220.

Sacks, H., E. A. Schegloff and G. Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation." Language **50**(4 (Part 1)): 696-735.

Schroeder, M. (2001). Emotional Speech Synthesis: A review. 7th European conference on Speech Communication and Technology (Eurospeech '01/Interspeech), Aalborg.

Schroeder, M. (2004). Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis, Saarland. **PhD**.

Schroeder, M., R. Cowie, E. Douglas-Cowie, M. Westerdijk and S. Gielen (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. EUROSPEECH-2001, Aalborg, Denmark.

Seneff, S. and J. Polifroni (2000). Dialogue Management in the Mercury Flight Reservation System. Satellite Dialogue Workshop, ANLP-NAACL, Seattle, USA.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg (1992). ToBI: A standard for labeling English prosody. International Conference on Spoken Language Processing, Banff, Canada.

Song, H. J., S. M. Ban and H. S. Kim (2009). Voice Activity Detection Using Singular Value Decomposition-based Filter. Interspeech 2009. Brighton, United Kingdom.

Sproat, R., M. Ostendorf and A. Hunt, Eds. (1999). The Need for Increased Speech Synthesis Research: Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis.

Stolcke, A., E. Shriberg, R. Bates, N. Coccaro, D. Jurafsky, R. Martin, M. Meteer, K. Ries and P. Taylor (1998). Dialog Act Modeling for Conversational Speech. AAAI 1998 Spring Symposium, Stanford University.

Suzuki, N. and Y. Katagiri (2003). Prosodic synchrony for error management in human-computer interaction. Error Handling in Spoken Dialogue Systems (EHSD-2003), Chateau d'Oex, Vaud, Switzerland.

Suzuki, N. and Y. Katagiri (2004). Alignment of human prosodic patterns for spoken dialogue systems. INTERSPEECH-2004, Jeju Island, Korea.

Suzuki, N. and Y. Katagiri (2005). Prosodic Alignment in Human-Computer Interaction Towards Social Mechanisms of Android Science, COGSCI 2005. Stresa, italy, Cognitive Science Society.

Swerts, M. and J. Terken (2002). " Dialogue and Prosody (editorial)." Speech communication(36): 1-3.

Syrdal, A., R. Bennett and S. Greenspan, Eds. (1994). Applied Speech Technology. Boca Raton, FL, CRC Press.

t'Hart, J., R. Collier and A. Cohe (1991). A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody. Cambridge, Cambridge University Press.

Tatham, M. and C. Morton (2005). Developments in speech synthesis. Chichester, England, Wiley.

Taylor, P. (1992). A Phonetic Model of English Intonation University of Edinburgh. **PhD**.

Taylor, P. (2000). "Analysis and Synthesis of Intonation using the Tilt model." The Journal of the Acoustical Society of America **107**(3): 1697-1714.

Titze, I. I. (1994). Summary Statement. Workshop on Acoustic Voice Analysis. Denver, Colorado (USA).

Trask, R. L. (1996). A Dictionary of Phonetics and Phonology. London,UK, Routlledge.

Traum, D. R. and J. F. Allen (1994). Discourse obligations in dialogue processing. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Las Cruces, New Mexico, Association for Computational Linguistics.

Vaughan, B., C. Cullen, S. Kousidis, W. Yi, C. McDonnell and D. Campbell (2006). The Use of Task Based Mood-Induction Procedures to Generate High Quality Emotional Assets. Information technology and Communications, IT&T, Carlow, Ireland.

Vaughan, B., Kousidis, S., Cullen, C., Wang, Yi. (2007). Task-Based Mood Induction Procedures for the Elicitation of Natural Emotional Responses. The 4th International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2007 Orlando, Florida.

Walker, J. and P. Murphy (2007). A Review of Glottal Waveform Analysis. Progress in Nonlinear Speech Processing. Berlin, Germany, Springer / Heidelberg**:** 1-21.

Wang, D. and S. Narayanan (2005). Speech Rate Estimation via Temporal Correlation and Selected Sub-Band Correlation. IEEE International Conference on Acoustics, Speech, and Signal

Processing, 2005 (ICASSP '05), Philadelphia, PA, USA.

Ward, A. and D. Litman (2007a). Automatically Measuring Lexical and Acoustic/Prosodic Convergence in Tutorial Dialog Corpora. SLaTE Workshop on Speech and Language Technology in Education, The Summit Inn, Farmington, Pennsylvania USA.

Ward, A. and D. Litman (2007b). Dialog Convergence and Learning. 13th International Conference on Artificial Intelligence in Education, Los Angeles, CA.

Ward, N. and S. Nakagawa (2004). "Automatic User-Adaptive Speaking Rate Selection." International Journal of Speech Technology(4): 259-268.

Warner, R. M. (2002). "Rhythms of Dialogue in Infancy: Comments on Jaffe, Beebe, Feldstein, Crown, and Jasnow (2001) " Journal of Psycholinguistic Research **31**(4): 409-420.

Weilhammer, K. and S. Rabold (2003). Durational aspects in Turn Taking. International Conference of Phonetic Sciences, Barcelona, Spain.

Werner, S. and E. Keller (1994). Prosodic Aspects of Speech. Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of tha Art, and Future Challenges. E. Keller. Chichester, John Wiley**:** 23-40.

Williams, J. D. and S. M. Witt (2004). "A Comparison of Dialog Strategies for Call Routing." International Journal of Speech Technology **7**(1): 9-24.

Wilson, M. and T. P. Wilson (2005). "An oscillator model of the timing of turn-taking." Psychonomic Bulletin and Review **12**(6): 957-968.

Woffit, R., N. M. Fraser, N. Gilbert and S. McGlashan (1997). Humans, Computers and Wizards: Human (Simulated) Computer Interaction. London, UK, Routledge.

Wright, H. F. (1999). Modelling prosodic and dialogue information for automatic speech recognition, Unviresity of Edingburgh. **PhD**.

Xiao, Z., E. Dellandrea, W. Dou and L. Chen (2005). Features extraction and selection for emotional speech classification. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), Como, Italy.

Xu, Y. (2005). "Speech melody as articulatorily implemented communicative functions." Speech Communication(46): 220-251.

Zellner, B. (1994). Pauses and the temporal structure of speech. Fundamentals of speech synthesis and speech recognition. E. Keller. Chichester, John Wiley**:** 41-62.

Zoltan-Ford, E. (1991). "How to get people to say and type what computers can understand." Int. J. Man-Mach. Stud. **34**(4): 527-547.