Synthesis of Listener Vocalizations

Towards Interactive Speech Synthesis

Dissertation

zur Erlangung des Grades des Doktors der Ingenieurwissenschaften
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Sathish Pammi

Saarbrücken, 2011

Tag des Kolloquiums: 16.01.2012

Dekan der Fakultät: Prof. Dr. Holger Hermanns

Mitglieder des Prüfungsausschusses: Prof. Dr. Dietrich Klakow
Prof. Dr. Hans Uszkoreit
Prof. Dr. Philipp Slusallek
Dr.-Ing. Marc Schröder

Kurze Zusammenfassung

Dialogsysteme nutzen zunehmend Hörer-Vokalisierungen, wie z.B. *a-ha* oder *mm-hm*, für natürliche Interaktion. Die Generierung von Hörer-Vokalisierungen ist eines der zentralen Ziele emotional gefärbter, konversationeller Sprachsynthese. Ein Erfolg in diesem Unterfangen hängt von den Antworten auf drei Fragen ab: Wo bzw. wann sollten Vokalisierungen synthetisiert werden? Welche Bedeutung sollte in den synthetisierten Vokalisierungen vermittelt werden? Und wie können angemessene Hörer-Vokalisierungen mit der intendierten Bedeutung realisiert werden? Diese Arbeit widmet sich der letztgenannten Frage.

Die Untersuchung erfolgt in drei Schritten: (i) *Korpuserstellung*; (ii) *Annotation*; und (iii) *Realisierung*. Der erste Schritt präsentiert eine Methode zur Sammlung natürlicher Hörer-Vokalisierungen von deutschen und britischen Profi-Schauspielern in einem Tonstudio. Im zweiten Schritt wird eine Methodologie zur Annotation von Hörer-Vokalisierungen erarbeitet, die sowohl Bedeutung als auch Verhalten (Form) umfasst. Der dritte Schritt schlägt ein Realisierungsverfahren vor, die Unit-Selection-Synthese mit Signalmodifikationstechniken kombiniert, um aus Nutzeranfragen angemessene Hörer-Vokalisierungen zu generieren. Schließlich werden *Natürlichkeit* und *Angemessenheit* synthetisierter Vokalisierungen mit Hilfe von Hörtests evaluiert. Die Methode wurde im Open-Source-Sprachsynthesesystem MARY implementiert und in den Sensitive Artificial Listener-Demonstrator im Projekt SEMAINE integriert.

Short Summary

Spoken and multi-modal dialogue systems start to use listener vocalizations, such as *uh-huh* and *mm-hm*, for natural interaction. Generation of listener vocalizations is one of the major objectives of emotionally colored conversational speech synthesis. Success in this endeavor depends on the answers to three questions: Where to synthesize a listener vocalization? What meaning should be conveyed through the synthesized vocalization? And, how to realize an appropriate listener vocalization with the intended meaning? This thesis addresses the latter question.

The investigation starts with proposing a three-stage approach: (i) *data collection*, (ii) *annotation*, and (iii) *realization*. The first stage presents a method to collect natural listener vocalizations from German and British English professional actors in a recording studio. In the second stage, we explore a methodology for annotating listener vocalizations – meaning and behavior (form) annotation. The third stage proposes a realization strategy that uses unit selection and signal modification techniques to generate appropriate listener vocalizations upon user requests. Finally, we evaluate *naturalness* and *appropriateness* of synthesized vocalizations using perception studies. The work is implemented in the open source MARY text-to-speech framework, and it is integrated into the SEMAINE project's Sensitive Artificial Listener (SAL) demonstrator.

This thesis is dedicated to my loving mother
for her love, endless support and encouragement.

Acknowledgements

Getting a doctorate degree has been my long-awaited dream. During this journey, many people helped me in several ways. I would like to thank each one of them.

First and foremost, I would like to thank our team leader and the SEMAINE project coordinator Marc Schröder for providing constant support in shaping my ideas, structuring my findings and discussing my results, always contributing with motivating and constructive comments. I am very grateful to him for being very supportive in difficult situations of my life as well. I feel deeply lucky as our paths met.

Many sincere thanks to the team members Marcela Charfuelan, Oytun Türk and Ingmar Steiner for helpful discussions and comments. I have really enjoyed working with all of you! My deep gratitude to Marcela Charfuelan and Holmer Hemsen for providing comments and suggestions to improve an earlier draft version of the thesis.

A special thanks to all of the SEMAINE project members for their valuable discussions in the project. During early stages of my PhD, elaborative discussions on foundations and basic concepts with Dirk Heylen, Catherine Pelachaud, and Elisabetta Bevaqua helped me a lot. Thanks to Roddy Cowie and Gary McKeown for their suggestions on perception studies.

My participation in Enterface'10 helped me to discuss my work extensively with the project members Dennis Reidsma, Khiet Truong, Daniel Neiberg, Iwan de Kok, and Herwin van Welbergen. Their suggestions helped me a lot to improve my thesis work. Thanks to Jürgen Trouvain for his critical comments and suggestions during the initial work.

I also thank Ivana Kruijff-Korbayova who kindly allowed me to spend more time on this thesis writeup during my contract with the ALIZ-E project; Stephan Busemann for his support in contract related matters.

I am very thankful to Kishore Prahallad and Sushma Bendre for inspiring me to choose research as my career path.

I would like to take this opportunity to express my deep gratitude to my family members for their love and encouragement, and for being supportive in hard times of life. I want to thank all my Indian friends here in Saarbrücken for the good times they have given me beyond academic life.

Last but not the least, I would like to thank my supervisor Hans Uszkoreit for giving me an opportunity to work in DFKI's LT lab and also for providing helpful guidance when I required the most.

Contents

List of Figures

Publications

The majority of the work presented in this thesis is based on papers published in the following conference proceedings.

SATHISH PAMMI AND MARC SCHRÖDER. “Evaluating the meaning of synthesized listener vocalizations". In *Proc. INTERSPEECH 2011*, Florence, Italy, 2011.

SATHISH PAMMI, MARC SCHRÖDER, AND MARCELA CHARFUELAN. “Multidimensional meaning annotation of listener vocalizations for synthesis". In *Proc. Workshop on Emotion and Computing*, Berlin, Germany, 2011.

SATHISH PAMMI, MARC SCHRÖDER, MARCELA CHARFUELAN, OYTUN TÜRK, AND INGMAR STEINER. “Synthesis of listener vocalisations with imposed intonation contours". In *Proc. Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*, Kyoto, Japan, 2010.

SATHISH PAMMI, MARCELA CHARFUELAN, AND MARC SCHRÖDER. “Multilingual Voice Creation Toolkit for the MARY TTS Platform". In *Proc.Language Resources and Evaluation (LREC)*, Valettea, Malta, 2010

BEVACQUA E., PAMMI S., HYNIEWSKA S., SCHRÖDER M., AND PELACHAUD C. “Multimodal backchannels for embodied conversational agents". In *Proc. Intelligent Virtual Agents*, pages 194–200, Philadelphia, USA, 2010. Springer.

SATHISH PAMMI AND MARC SCHRÖDER. “Annotating meaning of listener vocalizations for speech synthesis". In *Proc. Affective Computing & Intelligent Interaction*, Amsterdam, The Netherlands, 2009.

SATHISH PAMMI. “Synthesis of Nonverbal Listener Vocalizations". In *Proc. Doctorial consortium at Affective Computing & Intelligent Interaction*, Amsterdam, The Netherlands, 2009.

SATHISH PAMMI, MARCELA CHARFUELAN, AND MARC SCHRÖDER. “Quality control of automatic labelling using HMM-based synthesis". In *Proc. ICASSP 2009*, Taipei, Taiwan, 2009

Chapter 1

Introduction

With the ever-increasing role of computers in many areas of today's society, human-machine interaction has become an increasingly prominent part of our daily life. Machines and the ways people interact with them have changed dramatically in the past few decades. Traditionally, the human-machine interfaces have often been regarded as purely rational activity, in which emotions and social aspects are secondary. This view has been changing since the mid 90's when some studies (Langer 1992; Nass and Moon 2000) demonstrate that individuals mindlessly apply social rules and expectations to computers. People tend to interact with computers as if they were human-like. They unconsciously apply social rules even if they believe that such an attribution is not appropriate. Nowadays, humane-machine interfaces started considering emotions, social aspects and different intentions behind actual message to simulate human-like interactions.

Researchers who intend to make human-like human-machine interfaces started focusing mainly on building Embodied Conversational Agents (ECAs), a kind of intelligent humanoid graphical user interfaces, which can simulate human behavior like displaying facial expressions, moving head, performing gestures and making natural interaction with others like humans do in everyday life. To maintain natural and continuous interaction with humans, ECAs have to know how to react and respond based on interpreting what humans say and their non-verbal signals. If the interaction capabilities of ECAs are to become more human-like and they are to function in social settings, their design should support continuous interaction in which all partners in

an interaction perceive each other and express themselves continuously and in parallel (Thórisson 2002; Nijholt et al. 2008). In other words, human-like agents should be capable of simultaneous perception/interpretation and production of communicative behavior (Reidsma et al. 2011). They should be able to signal their attitude and attention while they are listening to their interaction partner (*active listening*), and be able to attend to their interaction partner while they are speaking (*attentive speaking*). However, many ECAs still remain immobile and fall silent while listening. *Active listening* is a structured form of listening and responding that focuses the attention on the speaker. The essential role of listening in natural interaction – to share mutual understanding with a dialogue partner – makes it a crucial issue in the development of ECAs.

This thesis is an investigation into the vocal part of *active listening* — the generation of listener vocalizations. In view of interactive speech synthesis, listener vocalizations play an important role in communicating listener intentions while the other person has the turn or is talking. The communicative intentions behind these vocalizations not only transmit messages like *I am listening* and *I am with you*, but also transmit their's affective states such as excited, bored, confused, surprised, and so on. Already a few attempts were made in this area of research; for example, the importance of affect bursts as a feedback in a conversation was investigated (Schröder, Heylen, and Poggi 2006) through listening tests, Ward and Tsukahara (2000) had developed some rules to generate back-channel responses in a conversation and investigated how to use low pitch regions as cues for back-channel responses. However, the analysis and identification of distinguishable types among back-channel vocalizations, their acoustic properties and the affective states behind them have to be studied as they are crucial to improve interactive speech synthesis.

1.1 Speech synthesis and the SEMAINE project

This thesis has been written in the context of the project SEMAINE (Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression)[1]. The project was concerned with developing Sensitive Artificial Listeners (SAL) (Schröder et al.

[1] http://www.semaine-project.eu

Figure 1.1: The four characters represent the four quadrants of arousal-valence space: Spike (top left), Poppy (top right), Obadiah (bottom left) and Prudence (bottom right). (Courtesy from (Schröder et al. 2009))

2008b; Heylen, Nijholt, and Poel 2007; Schröder et al. 2009; Douglas-Cowie et al. 2008) which are virtual dialogue partners based on audio-visual analysis and synthesis. Despite their very limited verbal understanding, they intend to engage the user in a conversation by paying attention to the user's emotions and nonverbal expressions. The system focuses on the "soft skills" that humans naturally use to keep a conversation alive. As part of the project, four psychologically different affective/personality types have been created to elicit different types of emotion – each employing individual dialogue strategies, and displaying uniquely different responsive reactions. Poppy is outgoing (extraverted) and optimistic; Spike is angry and argumentative; Prudence is pragmatic and practical; and Obadiah is gloomy and depressed. Figure 1.1 portrays these four SAL facial models.

The synthesized voices for the SAL characters are based on domain-oriented unit selection speech synthesis technology (Schweitzer et al. 2006; Schweitzer et al. 2003): they sound very natural within their domain (i.e., their usual types of utterances), but can say arbitrary text with a reduced quality. For each SAL character, a voice database has been recorded using a suitable, expressive actor's voice. This technological choice

produces the best currently available speech synthesis quality, but does not allow for an adaptation of prosodic properties of the speech; each SAL character will have its own expressive but constant speaking style. For non-verbal behavior, particularly listener behavior, the project planned to record vocalizations such as *laughter*, *sighs*, *hmmm*, *uh-huh*, etc. They become part of a behavior repository linking communicative functions to behaviors. The present thesis is concerned with this aspect of the voice: to provide the synthesis technology required to realize natural sounding listener vocalizations with the synthetic characters voice on the basis of intended meaning.

1.2 Motivation

Speech synthesis is used increasingly in interactive dialogue settings. Whereas early spoken dialogue systems adopted a ping-pong strategy for turn taking, newer spoken and multimodal dialogue systems attempt to model the computer's part of the dialogue in both the speaker and the listener role (Pfleger and Alexandersson 2004; Niewiadomski et al. 2009). That means the machine must emit signs of listening while the user is speaking: backchannels (Yngve 1970) or expressive feedback signals (Allwood, Nivre, and Ahlsén 1992). In multimodal dialogue systems, some of these signals can be visual, such as head nods, smiles, or raised eyebrows (Bevacqua et al. 2010); in the vocal channel, backchannel and feedback signals can be realized as listener vocalizations. Listener vocalizations like *mhm, right, yeah, uh-huh* are not only produced to make the interaction more natural but also to signal different meanings like *agreeing, interested, anger, etc.* (Pammi and Schröder 2009).

The generation of listener vocalizations is one of the major objectives of emotionally colored conversational speech synthesis. It includes different research questions like where to synthesize, what to synthesize and what kind of acoustic properties to realize in order to communicate different affective states in different situations. Therefore, success in this endeavor depends on the answers to three questions: Where to synthesize a listener vocalization? What meaning should be conveyed through the synthesized vocalization? And, how to realize an appropriate listener vocalization with the intended meaning? The major motivation of this thesis is to address the latter question. The first two questions are closely linked with dialogue structure and intension planning, which are outside the scope of the present work.

1.3 Thesis objectives

Although speech synthesis systems are providing high quality synthetic reading speech, more work is required to make speech synthesis is suitable for interactive settings. Interactive speech carries a great deal more information than just the verbal message. It can tell us the social stance, attitudinal and emotional states of dialogue partner. Needless to say that *expressive speaking* and *active listening* are basic 'soft skills' used to maintain interesting conversation. In order to integrate these skills into truly interactive multimodal dialogue systems, the current reading Text-to-Speech (TTS) systems have to be extended.

This thesis is a step towards the long-term objective of building socially plausible, reusable, interoperable interactive speech synthesis systems whose behavior is similar to a human interlocutor. The main objective of this research work is providing technological solutions to generate listener vocalizations by adding a new functionality to TTS that can synthesize listener vocalizations. In order to implement such functionality, the appropriateness to listener vocalizations should be investigated.

The intention behind this research is not only investigating listener vocalizations, but also building a system, which can be integrated into SEMAINE (Schröder et al. 2008b) project's multi-modal demonstration system, to synthesize listener vocalizations. The system has to be robust and it has to use standard representation like eXtensible Markup Language (XML) formats in view of intermodule communication. If the meaning and form of the vocalization is given in XML representation, the system has to generate an appropriate vocalization though recorded database has limited acoustic variety.

1.4 Research questions and thesis structure

The ability of human-like systems to generate listener vocalizations (i.e. the vocal part of *active listening*) is an important requirement for generating affective interaction. In order to give a system the ability to generate vocalizations that convey an intended meaning we need to address several research questions.

The main research question — Given a meaning to be conveyed through a synthesized vocalization, and the time to trigger it, what technological framework can generate an appropriate listener vocalization?

This question is addressed in this thesis by answering four concrete questions, that are explained in detail below.

RQ 1 — How to collect a database of natural listener vocalizations?

The first step in building corpus driven TTS systems is data collection. One of the most prevalent ways of collecting highly natural speech material is to record speech from an actor or speaker in a recording studio using pre-defined recording scripts. As listener vocalizations are produced naturally only in conversation, the corpus will not be natural if vocalizations is recorded with pre-defined recording scripts. The exercise to collect natural listener vocalizations is explained in Chapter 6.

RQ 2 — What kinds of meaning are conveyed through listener vocalizations? What is a suitable list of meaning and behavior descriptors to represent vocalizations?

In Chapter 7, we attempt to identify relevant categories of meaning for listener vocalizations in one of the corpus described in Chapter 6. This chapter describes our exploratory annotation approach to identify a suitable categorical description of the meaning conveyed in the vocalizations.

RQ 3 — How can appropriate listener vocalizations be identified that are appropriate for different meaning categories identified? Can any systematic patterns be found that relate this appropriateness to behavior properties?

The amount of work required for annotating meaning using multiple raters is huge. However, subjective meaning annotation is an essential task for generating appropriate vocalizations. Chapter 8 describes an approach that uses a multi-dimensional perception test to determine the appropriateness of listener vocalizations for a number of different meaning dimensions. The relevance of behavior properties like intonation and segmental form for the meaning perception of listener vocalizations is also investigated through a listening test.

RQ 4 — Given annotation of meaning and behavior (form) of listener vocalizations, how to realize an appropriate listener vocalization with a given intended meaning, in particular when a recorded vocalization does not exist?

An important limitation with simple unit selection approach to the synthesis of listener vocalizations is the fact that we can only generate the vocalizations that have been recorded. If we require additional vocalizations, such as an existing segmental form but with a meaning that had not been produced by the original speaker during the recording session, then the simple selection algorithm can only produce the vocalization most similar to the target – which may not be acceptable. Chapter 9 investigates a new extended unit-selection algorithm for selecting both candidate units and intonation contours, and for combining them. We finally evaluate the technological framework described to realize listener vocalization in Chapter 10. Two perception studies are conducted to evaluate the performance of synthesized listener vocalization for a given meaning.

The rest of the thesis is structured as follows. Chapter 3 discusses past related work on speech synthesis for interactive agents; furthermore the chapter presents relevant related work on *active listening* of ECAs. Chapter 2 looks closer into the influence of listener vocalizations on interaction. Chapter 4 provides background information on two relevant frameworks – MARY TTS and SEMAINE API. The investigation of the research starts with presenting a methodology in Chapter 5. Chapter 6 presents the speech corpus that will be used throughout this thesis. Chapter 7 concerns exploratory annotation to find meaning and behavior descriptors. In Chapter 8, the appropriateness of listener vocalizations is annotated with multiple raters. As Chapter 9 proposes an extended unit-selection algorithm to realize a listener vocalization, Chapter 10 evaluates the performance of the algorithm. Chapter 11 explains possible applications and usage of this work as a reflection of the efforts spent on this research. Finally, Chapter 12 concludes the research work involved this thesis.

Part I

Background

Chapter 2

Listener vocalisations

Listeners can talk (Yngve 1970; Gardner 2002). They bring their goals and desires into any conversation. Garcés Conejos and Bou Franch (2004) argue that the listener's metarepresentational, anticipatory and predicting abilities allow them to make inferences and produce listener responses in real time, at the speed with which the speaker is talking.

The listener role in a conversation is not an entirely passive one. According to Anderson and Lynch (1988), the listener is not just a "tape recorder"; listeners are selective, in terms of what they find most important or comprehensible or interesting in any particular message. Schegloff (1982a) argues that the discourse should be treated as an interactional achievement with a collaborative effort between the speaker and the other parties present. The essential information of the interactivity between the participants in a conversation is lost when we exclude the listener behavior.

The objective of this chapter is to provide some background information on listener responses. Section 2.1 gives some clarifications on the terminology used in this thesis. While Section 2.2 introduces the term *listener vocalizations*, Section 2.3 discusses some background studies on these vocalizations. In Section 2.4, we discuss the functions of the vocal listener behavior. Section 2.5 describes the meanings conveyed by listener vocalizations. Section 2.6 provides some information on the characteristics of the vocalizations. Finally, Section 2.7 summarizes the content of this chapter.

2.1 Terminology

2.1.1 Intentions and communication

Intentions are mental states that humans develop rationally in order to enable the fulfillment of their desires, given their beliefs (Bratman 1987; Bratman 1990). According to pragmatics (Haugh 2008), communication involves speakers expressing their intentions, and listeners attributing intentions to those speakers. The success of the communication indeed depends on how exactly listeners interpret speakers' original intentions.

Sperber and Wilson (1995) in their relevance theory deconstructed intentions into two types: (i) Informative intention – the intention to manifest or more manifest to the audience a set of assumptions. (ii) Communicative intention – the intention to make it mutually manifest to the audience and the communicator that the communicator has this informative intention. In simple words, a speaker's informative intention is to make listeners believe certain things, whereas the communicative intention is to have his informative intentions recognized. A successful communicative intention leads the listener to understand the speaker's intended meaning.

Human communication is crucially dependent on the existence of communicative intentions, which exist in the mind of a speaker, and about which listeners make inferences. Levinson (1983) defined communication as follows:

> "... communication consists of the 'sender' intending to cause the 'receiver' to think or do something, just by getting the 'receiver' to recognize that the 'sender' is trying to cause that thought or action. So communication is a complex kind of intention that is achieved or satisfied just by being recognized. In the process of communication, the 'sender's' communicative intention becomes mutual knowledge to 'sender' (S) and 'receiver' (H), i.e. S knows that H knows that S knows that H knows (and so ad infinitum) that S has this particular intention. Attaining this state of mutual knowledge of a communicative intention is to have successfully communicated." (Levinson 1983, p.16)

2.1.2 Floor and turn

According to conversational analysis, *floor* is the right to speak in an interaction between two or more persons. People take *turns* in order to get control of the right to speak.

> "When two people are engaged in conversation, they generally take turns. First one person holds the floor, then the other. The passing of turn is nearly the most obvious aspect of conversation. .... a person engages in different activities when he has the turn than he doesn't have it. When he has the turn he engages primarily in speaking activities and when he doesn't have it he engages primarily in listening activities." (Yngve 1970, p.567)

Drummond and Hopper (1993) further explained the difference between the floor and the turn as follows:

> "In the front channel, each speaker's turn also alternates the floor. But when one speaker gains the floor for an extended topic or story, the recipient may take brief turns with out gaining the floor." (Drummond and Hopper 1993, p.159)

Turn-taking can be cooperative or competitive (French and Local 1983). Speakers may cooperate and share the floor equally; they may compete for keeping the floor and preventing others from getting it.

2.1.3 Function, meaning, and behavior

In the literature, the terms *function* and *meaning* are often used interchangeably when talking about listener responses. It seems difficult to draw a clear line between functions and meanings. *Function* seems to be a more general term than *meaning*.

The meaning can be classified broadly into two categories, though they are closely related, based on the way we study listener responses: semantic meaning and pragmatic meaning. *Semantic meaning* represents the first level of meaning, the significant meaning, i.e. what the listener response "says" in general; *Pragmatic meaning* represents the second level meaning, the situational meaning, i.e. what does the listener response really mean in the actual situation/context.

Behavior (form)

Behavior usually refers to the actions of a person. In multimodal interaction, the term *behavior* is used to talk about visual signals such as gestures, facial expressions, body movements. In this thesis, we adopted this term to refer to vocal listener behavior, the form, such as the segmental form and prosody of listener vocalizations.

2.2 What are listener vocalizations?

While listening we do not stand silent, we produce vocalizations such as *yeah* and *uh-huh*, move around, nod and gaze etc. All those audible and visible acts have a meaning and regulate the conversation.

Terms	**Used by**
Signals of continued attention	Fries (1952)
Concurrent feedback	Krauss and Weinheimer (1966)
Accompaniment signals	Kendon (1967)
Listener responses	Dittmann and Llewellyn (1968)
Backchannels or Backchannel responses	Yngve (1970); Duncan (1974); Ward and Tsukahara (2000)
Acknowledgment acts	Sinclair and Coulthard (1975)
Minimal responses	Fishman (1978)
Hearer signals	Bublitz (1988)
Receipt tokens	Atkinson (1992)
Affirmative responses	Hirschman (1994)
Reactive tokens	Clancy et al. (1996)
Minimal feedback	Holmes (1997)
Response tokens	Gardner (2002)

Table 2.1: Terminology used in the literature to represent listener vocalizations

A lot of research has addressed listener behavior over past four decades. Throughout the literature, the term *back channel* has been widely used to talk about listener

behavior despite a little controversy over the usage of the term (Fujimoto 2009). However, a vast terminology is being used by several researchers to talk about listener behavior (as shown in Table 2.1).

The definitions of the above terms vary and each choice reflects a very specific methodological approach to the phenomenon studied. Among the list of terms used in the literature, however, the term *listener response* seems to be easily comprehensible. Moreover, it can be used for multimodal listener responses as well. This thesis uses the term *listener vocalization* to refer to the vocal part of the listener response.

The operational definition of *listener vocalizations* used in this research work is the following:

> "brief vocal responses ... by the nominal listener, which do not constitute an attempt to take the conversational floor." (Bilous and Krauss 1988, p.186)

2.3 Some pragmatic perspectives

Fries (1952) seems to have been the first to talk about listener vocalizations in English conversation, and he noted that vocalizations such as *uh huh*, *yeah*, *I see* are used to show continued attention. Yngve (1970) coined the term *back channel* to denote listener vocalizations such as *yes* and *uh-huh*.

> "In fact, both the person who has the turn and his partner are simultaneously engaged in both speaking and listening. This is because of the existence of what I call the back channel, over which the person who has the turn receives short messages such as "yes", and "uh-huh" without relinquishing the turn." (Yngve 1970, p.568)

Duncan (1974) outlined five types of listener behavior comprising of vocal, verbal and gestural forms: (i) Readily identified, verbalized signals such as *yeah*, *mhm*, *right*; (ii) Sentence completions; (iii) Requests for clarifications; (iv) Brief statements; (v) Head nodes and shakes.

Type	Definition	Display	Examples
Backchannels	a non-lexical vocalic form which serves as a 'continuer'	interest, or claim of understanding	*hm, huh, oh, mhm* and *uh huh*
Reactive expressions	a short non-floor-taking lexical phrase	feedback or assessment	*yeah, sure, exactly, shit, good* and *wow*
Collaborative finish	an utterance produced by the non-primary speaker to finish a previous speakers utterance	collaboration regarding management of conversational Floor	*hm, hmm* and *mhm*
Repetitions	a portion of the speech of the primary speaker repeated by a non-primary speaker	clarifications	
Resumptive openers	a non-lexical vocalic form which is followed by a full turn	acknowledge the prior turn	

Table 2.2: The functions of non-primary speaker (Summarized from Clancy et al. 1996)

Types	Discourse function	Examples
continuers	keep the floor open for the current speaker to continue speaking	*mh mh*, *uh huh*
acknowledgments	claim agreement or understanding of speaker's turn	*mh*, *yeah*
newsmarkers	mark what the speaker has said as newsworthy in some way	*really?, oh!, right!*
change of activity tokens	mark the transition for a new activity or topic	*okay*, *alright*
assessments	evaluate the talk of the current speaker	*great*, *wow*
clarifications	check to make sure s/he has heard correctly	*who?*, *huh?*
interest or attentive signals	display interest and engagement in what the speaker has said	*gosh*
collaborative finish	the listener finishes the speaker's utterance	
emotional response	expresses an emotional reaction to the speaker	*(laughter), (sigh)*

Table 2.3: Types of listener vocalizations and their discourse functions with archetypical examples (Gardner 2002)

Goodwin (1986) shows that some listener responses systematically occur within phrases overlapping the interlocutor's speech, whereas others can occur in the brief pauses between the interlocutor's pauses. He differentiated them between two types: continuers and assessments. According to him, continuers are "actions displaying recipient's understanding that an extended turn at talk is in progress but not yet complete", such as head nods and *uh-huh*; whereas, listener's assessments comment on the specifics of what is being, and has been, said by interlocutor, such as *oh wow, good* and *beautiful*.

Bavelas, Coates, and Johnson (2002) describe listeners as co-narrators from his narrative storytelling experiments. He noted two types of listener responses: generic responses and specific responses. Generic responses, such as nods, *yeah*, and *mhm*, convey attentiveness and understanding but not specifically related to what the listener was saying at the moment. Specific responses, such as winces, frowns, or supplying words, were tightly connected to that precise moment in the speaker's narrative.

Clancy et al. (1996), in their study which has a goal to examine communicative strategies of non-primary speakers, distinguished between five types of listener responses (see Table 2.2) : Backchannels, Reactive Expressions, Collaborative Finishes, Repetitions, and Resumptive Openers. Gardner (2002) summarized types of listener vocalizations used in a conversation from previous studies (see Table 2.3).

2.3.1 Influence of segmental form

This section summarizes eight frequently discussed listener vocalizations in the literature. They are *mm-hm, uh-huh, yeah, mm, oh, right, okay* and *alright*.

Jefferson (1983) compared listener vocalizations *uh-huh* and *mm-hm* with a third item, *yeah*. She argued that *yeah* displays speakership incipiency, whereas *mm hm* displays passive recipiency. This argumentation was later confirmed by Drummond and Hopper (1993), in their words, "*yeah* displays speakership incipiency, or actions to end recipient status and to share or take the floor". They used the term *acknowledgment tokens* for listener responses and tried to distinguish these tokens based on degree of speakership incipiency. They argue that continuers, such as *mm hm* and *uh huh*, show low degree of speakership incipiency when compared to acknowledgments, such as *yeah* and *mm*. A large section of previous studies confirms that *mm-hm* and *uh-huh*

are prototypical continuers, both of them can be used as the same way in order to take passive incipiency.

The tokens *oh* and *right* were identified as newsmarkers which mark the prior turn as newsworthy. Schiffrin (1988) claimed that the token *oh* is used to mark transition between information states of the interlocutor's speech. Gardner (2002) noted that the most common tokens *okay* and *alright* are used for marking a change of activity or a change of topic. Moreover, *alright* showed stronger shift in change of activity than *okay*.

2.3.2 Influence of prosody

The prosody of listener vocalizations influences the meaning and discourse level functions. Two previous studies that support this argumentation are described in this section.

Ward (2006) has worked on non-lexical "conversational grunts" of American English such as *oh, um* and *uh-huh*, and he attempted to correlate the sound and meaning at a lower level, i.e. "sound symbolism" in his words, see Table 2.4.

sound	meaning
syllabification	lack of desire to talk
duration	amount of thought
loudness	confidence importance
pitch downslope/upslope	degree of understanding / lack thereof
pitch height	degree of interest
creaky voice	claiming authority

Table 2.4: Meanings attributed to some prosodic features (Ward 2006)

Gardner (2002) investigated the influence of intonation on the token *mm*. He explores the effect that three types of intonational contours have on *mm* as listener vocalization: the falling contour, the fall-rising contour and the rise-falling contour. He noted as follows:

- *mm* (fall-rise intonation) is used as continuers such as *mm-hm* and *uh-huh*

- *mm* (falling intonation) is used as acknowledgments such as *yeah*
- *mm* (rise-fall intonation) is used as assessments such as *great* and *wow*

Yeah and *mm* are two peculiar vocalizations (Gardner 2002), they can be used as both continuers and acknowledgement tokens in a conversation. With slightly falling intonation, on the one hand, they act as acknowledgement tokens; on the other hand, they act as continuers with rising intonation.

2.3.3 Influence of culture and gender

Listener behavior is somewhat widely studied conversational phenomena which has been claimed to show cultural and gender related differences. Some of them are discussed in this section.

Stubbe (1998) investigated cultural differences in New Zealand English conversations of two ethnic groups Maori and Pakeha. She noted statistically significant difference between the two groups in the amount of feedback produced in conversations. She mainly observed mismatches in three different areas of the listening behavior: the use and the interpretation of pauses and silences, the appropriate amount of feedback, and the degree to which the supporting feedback is overt or indirect.

Gardner (2002) wrote that the most frequent use of the token *right* in British and Australian English is as a newsmarker, whereas Americans' mostly use it as an agreement marker. Japanese speakers are more likely to produce semantically empty listener responses like *mm-hm* and *uh-huh* whereas Americans preferred to use contentful ones like *yeah* (Maynard 1997). Tottie (1991) noted that American English conversations contain 16 listener responses in a minute whereas British English conversations have only 5 listener responses per minute. In his review on listener responses, Xudong (2009, p.118) wrote: "...people from different cultural groups may use listener responses differently in terms of their frequency of use, their placement in the conversational context and in terms of the different types of listener responses."

Many studies claim that women use more listener responses than men do both in mixed-gender conversation and in same-gender conversation. For example, Stubbe (1998) found a higher proportion of overtly supportive feedback was observed in women than men irrespective of ethnic group. Feke (2003) investigated gender differences in listener behavior based on conversations from native speakers of English

and Spanish. The observations of the study are as follows: (i) Both native speakers use more responsive behavior in mixed-gender conversations when compared to same-gender conversations. (ii) Females produce more responsive behavior than do males in mixed-gender conversations. Leet-Pellegrini (1980) reported that women used more assessments such as *yeah, right* and *uh-huh* than men did.

2.4 Functions of listener responses

Some of the primary functions of listener responses are mentioned in the literature as follows:

- signal attention (Fries 1952)
- show involvement (Dittmann and Llewellyn 1968)
- acknowledge the ongoing telling (Drummond and Hopper 1993)
- indicate agreement (Kendon 1967)
- express disagreement (Brunner 1979)

Allwood, Nivre, and Ahlsén (1992), in their studies on listener responses, argue that listener responses in conversation are used to fulfill broadly three functionalities: *feedback*, *turn management* and *sequencing*. According to them, listener responses enable the participants of a conversation to unobtrusively exchange information about four basic communicative functions: contact, perception, understanding and attitudinal reactions. The attitudinal reactions are further divided into accept, reject, belief, agreement, surprise, etc.

Manusov and Trees (2002) distinguishes the nonverbal behavior of the listener between attitudes/affect, feedback and backchannels. They further characterized listener responses on 5 scales: (i) positive/negative affect; (ii) certainty/uncertainty; (iii) confusion/understanding (iv) agreement/disagreement; (v) belief/disbelief. Heinz (2003) used six categories to represent listener responses: supports, exclamations, exclamatory questions, awareness/wonder, hesitation and negation.

Classifying the functions of listener vocalizations is difficult because their semantic and pragmatic meaning seems somewhat complex and involving several dimensions. However, Garcés Conejos and Bou Franch (2004) described that listener behavior mainly includes three types of functions: they are *cognitive*, *social* and *discourse regulatory functions*.

2.4.1 Cognitive functions

The aim of cognitive functions of listener responses in a conversation is to let the interlocutor know whether the listener is processing new information and making appropriate inferences while the speaker is talking.

> "The cognitive function of listener responses is to signal whether the assumption in the speaker's utterance has been combined with other information known to the listener and, following the path of least effort, has achieved relevance: the listener, then, lets the speaker know of the state of the interpretation process." (Garcés Conejos and Bou Franch 2004, p.30)

The cognitive functions involve, 'theory of mind' (Baron-Cohen 1988), that is, the ability to communicate mental states such as beliefs and desires to others.

2.4.2 Social functions

Listener vocalizations have a social function in "phatic communion"[1] (Richards 1983), and they express an emotional or attitudinal stance toward the interlocutor's utterance. The nonverbal behavior of listener vocalizations is a key to communicate social functions behind the responses. Social signals include (dis-)interest, (dis-)agreement, anticipation, empathy, hostility, and any other attitude towards others.

> "The social function of listener responses has been related to their signaling (lack of) involvement, affect and/or interest." (Garcés Conejos and Bou Franch 2004, p.31)

[1] According to Lyons (1968), 'phatic communion' is a spoken communication that serves to establish and maintain a feeling of social solidarity and well-being. It is more about sharing feelings or establishing an atmosphere of sociability rather than communicating facts and ideas.

Brown and Levinson (1987) describe listener responses such as acknowledgements and repetitions as satisfying the interlocutor's positive 'face needs'. The responses may function variably to express solidarity according to the different face needs of the participants. These responses can also be non-supportive or challenging.

2.4.3 Discourse regulatory functions

Discourse regulation is about distributing turns at talk in ongoing discourse. The listener analyses the interlocutor's speech in order to identify transition-relevant places for turnshifts. The discourse regulatory function deals with contextually appropriate selection of speaking-hearing roles (Garcés Conejos and Bou Franch 2004), where cultural expectations and participant's mental (meta)representations play a significant role to shape the discourse.

Kasper (1989) classified the discourse regulatory functions into four types: "uptaking", "turn-taking", "turn-keeping" and "turn-giving". Among them, the first two are relevant to the state of listener behavior. The listener's 'uptaking' indicates that s/he follows the interlocutor's contribution without making any claims for turn-shift, whereas 'turn-taking' may takes place when the listener claims for turn-shift at a transition relevance place. However, the interlocutor sometimes attempts to keep the floor when the speech flow is disrupted.

2.5 What do listener vocalizations convey?

Listeners not only communicate their affective and epistemic states to the dialog partner, but also continuously update them according to the interlocutor's speech. *Gestural act* can be defined as "a planned or unplanned gesture meant to indicate its own state or to change the information state of the receiver" (Versloot 2005). We consider that the listener's vocal behavior is part of his/her gestural acts. The states behind these acts include *affective states*, *epistemic states* and *turn management cues*. This section discuss these three types.

2.5.1 Affective states

Emotions are part of human life. With the help of empirical studies, Cowie, Sussman, and Ben-Ze'ev (2011) found that there are almost no instances where people report their state as completely unemotional, though full-blown emotions are quite rare. Listener's emotions are about the feelings regarding the content of interlocutor's speech.

Affective states represent all the emotional expressions of the listener, planned or unplanned. Listener vocalizations convey the listener's affective states such as *anger, amusement, sadness* and *so on*. Stubbe (1998) described a section of listener vocalizations as Supportive minimal responses(SMRs) by which listeners signal affectively positive meanings such as sympathy, interest or surprise.

2.5.2 Epistemic states

The concept of epistemic states was highlighted by Baron-Cohen et al. (2004), also called mental states or cognitive states. These states represent higher level attributes of mind than basic emotions, i.e. "state of mind". The ability to attribute epistemic states with content to others has been called a "theory of mind" (Baron-Cohen 1988). While the concepts of physical world can be called "primary representations", the concepts of other people's mental states are representations of other representations, i.e. "secondary representations" or "meta-representations". According to Baron-Cohen (1988), the meta-representations are essential for social skills. Listener responses plays a crucial role in communicating listener's meta-representations to the interlocutor.

The epistemic states behind these vocalizations not only show involvement messages such as *I am listening* and *I am with you*, but also convey the listener's epistemic states such as *attention, interest*, and/or *understanding* (Fries 1952; Kendon 1967).

2.5.3 Turn management cues

Many instances of listener's behavior include turn-taking cues. Schegloff (1982b) noted that sometimes listeners have the chance to begin to speak but do not and instead use vocalizations such as *uh-huh* and *yeah* to allow the interlocutor to continue the turn. Drummond and Hopper (1993) attempted to classify listener vocalizations according to the degree of speakership recipiency. *Continuers* shows no degree of speakership

recipiency, whereas *acknoledgements tokens* contains some degree of speakership recipiency.

2.6 Characteristics of listener vocalizations

Listener behavior is a complex discourse feature. The characteristics of this feature are determined by “a very fluid relationship between form and function" (Stubbe 1998). This section describes two main characteristics of listener vocalizations.

2.6.1 Multifunctional nature

Many listener vocalizations are multifunctional in nature. They include several functions of *cognitive*, *social* and *discourse regulatory functions*. They communicate several meanings at the same time. The “sound-meaning" relation among listener vocalizations has fairly complex structure.

Ward (2006) argues that each of these utterances means many things at many levels. Schegloff (1982b) observed the multifunctioning behavior in the token *yeah*: it marks, on the one hand, acknowledgment and confirm understanding; it may also express agreement, on the other hand. In this way, social actions are coordinated and fine-tuned on several levels simultaneously. McCarthy (2003) also noted the multifunctional behavior in listener vocalizations.

2.6.2 Appropriateness of behavior

In order to maintain a smooth conversation, listeners are expected to provide appropriate feedback to communicate their involvement in a conversation. They use appropriate listener vocalizations to reflect their affective and epistemic states.

- **Semantically appropriate:** Listeners convey their internal states through their vocal responses. Such responses are semantically appropriate with the listener’s cognitive stance such as *interest level* and *attention*.
- **Contextually appropriate:** In order to influence and regulate the discourse, listeners not only reflect their own states but also consider the context of their interlocutor’s speech.

- **Culturally appropriate:** Many studies observed that the listeners obey their cultural expectations. They choose appropriate vocalizations in order to meet the expectations.

2.7 Summary

Listener vocalizations are manifestations of a listener's intentions and his/her metarepresentations towards others. As these intentions contain several epistemic-affective states and some discourse level aspects, it is fairly complex to understand the nature of listener behavior completely. Many research studies reviewed in this chapter investigated listener behavior from the view points of cognitive, social, and pragmatic theories, though each of the studies is limited to a particular dimension. This chapter also discussed the meanings conveyed by the vocalizations based on previous research studies. Finally, we discussed the characteristics of listener vocalizations such as multifunctionality and appropriateness.

Chapter 3

Speech synthesis and interactive agents

Present day research on interactive agents has increased its focus on different spoken dialogue settings. Embodied conversational agents (ECAs) are demanding natural, spontaneous, interactive synthetic speech. Several recent investigations are aimed to reach such demands. Although the current technology met some of them such as high quality reading synthetic speech, there is a long way to travel in order to reach several objectives such as high quality interactive and spontaneous synthetic speech. This chapter provides some background information on recent work in emotional and conversational speech synthesis. With a primary concern on listening behavior, we also discuss interactiveness in several interactive agents or virtual humans.

This chapter starts with a brief introduction on the state-of-art of speech synthesis technologies (see Section 3.1): unit selection approaches and statistical parametric approaches based on Hidden Markov Models (HMMs). Section 3.2 reviews some interesting investigations on spontaneous synthetic speech techniques such as expressive and conversation-like speech synthesis. We discuss the need for incorporating *attentive speaking* and *active listening* skills into ECAs. In Section 3.4, we review several ECAs, which are able to realize listening behavior, developed in the literature. Section 3.5 summarizes the chapter.

3.1 Speech synthesis

Speech synthesis is the process of converting text into a speech signal. The objective of Text-to-Speech (TTS) synthesis is to convert any arbitrary input text to intelligible and natural sounding speech so as to transmit information from a computer to a human. This section provides a very brief overview on the current popular speech synthesis techniques: unit-selection based and HMM-based TTS systems.

The unit-selection algorithms are well known for natural sounding speech synthesis. In contrast, HMM-based parametric speech synthesis is popular for intelligible systems. In addition, HMM-based speech synthesis is flexible due to its parametric modeling process which can allow changing voice characteristics, emotions, and speaking styles.

3.1.1 Unit selection based approach

The unit-selection based approaches are based on: the selection of appropriate candidate units, which are close to the intended *target*, from a database of natural speech; and a appropriate combination of the selected units in order to achieve good quality speech.

The unit-selection algorithm plays a key role in identifying which of the available candidate units are appropriate for the target of intended speech to be synthesized. Hunt and Black (1996) presented a cost-based selection algorithm which includes two types of costs: (i) *target cost* – it defines how well a candidate unit from the database matches the target unit; (ii) *concatenation cost* – it defines how well two selected units combine.

Black and Taylor (1997) went another step further and proposed a decision tree based pre-clustering of candidate units which allows the runtime synthesis system to find similar units by asking questions on phonetic and prosodic contexts, as shown in Figure 3.1.

3.1.2 HMM-based approach

HMM-based speech synthesis comprises four steps: (i) *Parameter extraction* – the extraction of parameters from the utterance database; (ii) *Model training* – The training

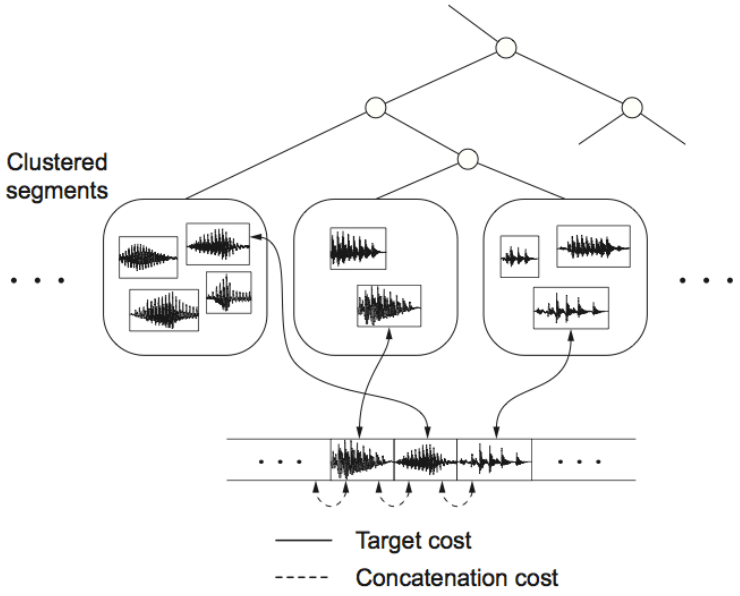


Figure 3.1: Clustering units in selection algorithm (Zen, Tokuda, and Black 2009)

of the HMMs that model the extracted parameters by taking contextual factors into account (iii) *Parameter generation* – the parameter generation from the given text using the trained HMM models; (iv) *Vocoding* – the waveform generation using a suitable vocoder. The first two steps and the latter two steps are also called “training" and “synthesis" stages respectively (Black, Zen, and Tokuda 2007). Among these four steps, a simple “re-synthesis" (i.e. copy-synthesis) includes *parameter extraction* and *vocoding* steps only. Figure 3.2 shows an overview of HMM-based speech synthesis.

Parameter extraction

In this stage, spectrum and excitation parameters are extracted from a speech database. The spectral parameters include mel-cepstral coefficients and their dynamic features, whereas excitation parameters contain strengths, magnitudes, log F0 and their dynamic features.

Model training

This stage includes the training phase of spectrum, pitch and duration HMMs. This process is very similar to that for speech recognition. In this training procedure, both

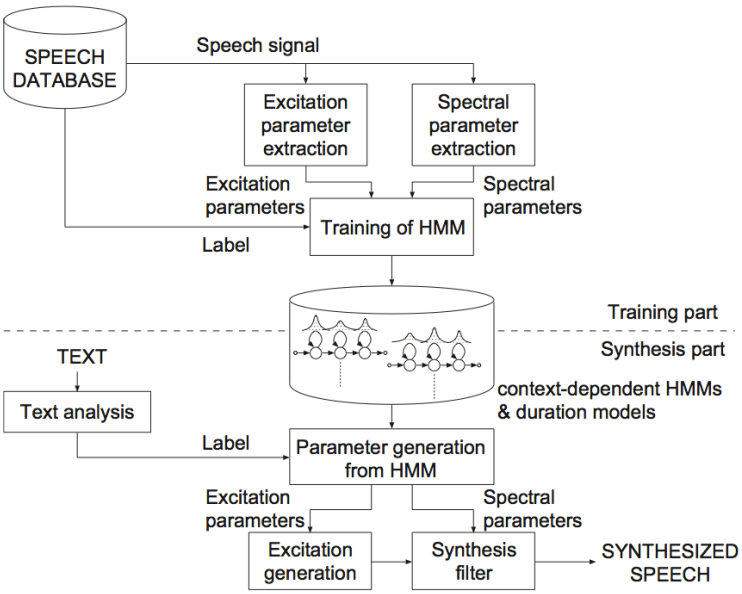


Figure 3.2: Overview of HMM-based speech synthesis (Black, Zen, and Tokuda 2007)

spectrum and excitation parameters are modeled by a set of context-dependent HMMs (CD-HMMs). By taking phonetic, linguistic, and prosodic contexts into account, context-dependent HMMs are trained to model the extracted mel-cepstral coefficients and their dynamic features. Similarly, Log F0 can be modeled by a hidden Markov model based on multi-space probability distribution (MSD-HMM) (Yoshimura et al. 1999), and State duration densities can be modeled by single Gaussian distributions.

Due to the exponential growth in the combinations of contextual factors, several researchers (e.g. Odell 1995; Miyazaki et al. 1998) used a decision-tree based context clustering algorithm. The distributions for spectrum, pitch and state duration are clustered independently since each of them has its own influential contextual factors.

Parameter generation

Parameter generation is a runtime step required to generate parameters from input text. First, the given input text is converted into a context-dependent label sequence; with the help of such sequence, an utterance HMM is constructed by concatenating the CD-HMMs. From the utterance HMM, the parameter generation algorithm (Tokuda

et al. 2000) generates the sequences of not only spectral parameters such as mel-cepstral coefficients and their dynamic parameters, but also excitation parameters such as strengths, magnitudes, log F0 and their dynamic parameters.

Vocoding

This step can be viewed as a inverse step to the parameter extraction. The mel-cepstral analysis technique that is used in the parameter extraction phase enables speech to be re-synthesized from the spectral parameters using the MLSA (Mel Log Spectrum Approximation) filter (Imai 1983). This vocoder uses generated spectral and excitation parameters in order to realize synthetic speech.

3.2 Spontaneous synthetic speech

Expressivity and spontaneous nature are the current challenges for synthetic speech. The current technologies that are used in interactive agents will require more conversational like synthetic voices. Such voices must simulate the way people talk instead the way people read. Current research is focussing on emotionally colored conversational speech synthetic systems that include disfluencies, filled pauses, hesitations, affect bursts, listener vocalizations etc. This section briefs about some state-of-art studies in this direction.

3.2.1 Expressive speech synthesis

Schröder (2009) reviewed several recent studies on emotional speech synthesis and categorically divides the available approaches into “explicit", “play-back", and “implicit" models. This section adopts the classification and briefs these approaches.

Explicit models

Explicit speech synthetic models aim for general purpose systems that are able to express several emotional states based on the link between emotions and their prosodic realizations. The key for these models is to define emotion specific global prosodic settings, such as F0 level and range, speech tempo and loudness.

Zovato et al. (2004) investigated signal modification techniques such as PSOLA (Pitch Synchronous and Overlap Add) to impose emotional prosody rules on selected units. This approach facilitates explicit modeling, however, it has the disadvantage of creating audible distortions for larger modifications. Schröder (2006) proposed a set of emotional prosody rules to express gradual emotions in synthetic speech based on a literature review. The hand-crafted prosody rules are implemented to reflect the prosodic settings on diphone voices available in the MARY[1] text-to-speech synthesis system. The results of the work confirmed that the prosody rules are able to express a *continuum* of activation.

Though the explicit models achieved good recognition rates for target emotion, their synthesized speech is unnatural and exaggerated due to the hand-crafted prosody rules. As a result, they are not popular for widespread usage.

Playback models

'Play-back' models are about reproduction of expressivity from the speech database recordings. The expressivity is not explicitly modeled but rather selected from the corpus. Here, the speaker is asked to make separate recordings for different emotions; a given emotional synthetic speech can be generated by selecting units only from the corresponding subset of the recordings (Iida and Campbell 2003). This methodology is best-suited for limited or specialized applications.

Fernandez and Ramabhadran (2007) attempted to incorporate expressivity into symbolic target; that allows the model to strictly enforce the selection of units from given emotion. In addition, they used a similarity cost matrix for emotions, which improves flexibility in choosing units in the case of unavailability of intended style specific units; this allows the model to choose units from similar emotions. If the intended speaking style is *sad*, for example, it would be possible to penalize *anger* units more than *neutral* units using the similarity matrix.

Steiner et al. (2010) made an attempt to compare the performance between symbolic and acoustics-based style control for expressive unit selection. Their experimental corpus consists of a relatively large body of neutrally spoken speech material and

[1] http://mary.dfki.de

it includes, additionally, four expressive speaking styles; *cheerful, depressed, aggressive* and *poker* (cool, laid back) speaking styles. All of the recordings were made by a single professional German actor. While increasing weight of the symbolic feature in unit selection algorithm, the majority of synthesized units' distribution were found in the specified target style; as a result, the spectral distance of synthetic speech became closer to the original speaker's stylized speech (gold standard). In the case of acoustics-based style control, they used open quotient gradient (OQG), F0 and duration as acoustic targets. However, this type of control showed only a partial success because they used limited acoustic features as targets. The evaluation results are based on objective measures such as spectral distances and distribution of selected units for synthesis, but the relevance of perceptual evaluation remains to be investigated. They also claimed that further work is needed in this direction of acoustics-based style control.

Yamagishi et al. (2003) experimented with using HMM-based parametric speech synthesis models in order to train four different speaking styles: "reading", "rough", "joyful" and "sad". On the one hand, they trained fully separate HMM-based voices; on the other hand, they created a combined voice in which style was part of context description. In both kinds of voices, they showed that the parametric synthesis technology can reproduce style, through perception studies. Yamagishi et al. (2007) further attempted to use *adaptation* rather than training in order to reproduce a speaker's speaking style in synthetic speech. They created two voices: the first one was built by *training* HMM models with 453 sentences from the target speaker; the second one was created by *adapting* spectrum, F0 and duration of an average voice to 100 sentences from the same speaker. Interestingly, they showed that synthetic speech of the *adapted* voice is more similar to the original speaker's speech when compared to synthetic speech of the *trained* voice.

Implicit models

'Implicit' models are corpus driven approaches where a target is predicted by one model and such target is realized by another model; but these two models should be created either by analyzing the same corpus or by learning models from that corpus. In other words, implicit models formulate or learn rules from the same corpus used for

building expressive voices, opposing explicit hand-crafted rules. These approaches are being used by present researchers in order to get a balance between the flexibility in expression and the quality of the state-of-art systems.

With the objective of incorporating *emphasis* into the synthetic speech, Fernandez and Ramabhadran (2007) explored a boot-strap training mechanism to emphasize words in recording corpus. Among 10,000 sentences in their corpus, the words of around 10% of the sentences were manually labelled with *emphasis*; whereas, remaining words in the corpus were labelled with statistical models that are trained with the manual labels. They used different symbolic labels for manually vs. automatically labelled emphasis in order to slightly favor manually labelled units over automatically labelled units during synthesis.

Inanoglu and Young (2009) investigated data-driven emotion conversion strategies. Their objective was to convert neutral speech of a speaker into an emotional speech (anger, surprise and sadness) of the same speaker. They proposed a three stage strategy for emotion conversion from neutral to a specific style (see Figure 3.3). The first stage involves spectral conversion, applying a GMM-based linear transformation method to a sequence of vocal tract features extracted from LPC-analysis, to change the neutral voice quality to that of a target emotion; the transformed frames were combined with Overlap-Add (OLA) synthesis. In the second stage, a set of regression trees predict phone durations of the target emotion by using neutral phone durations and their linguistic context features as input; then, a duration conversion mechanism is applied to the first stage synthesized waveform using TD-PSOLA (Time Domain Pitch synchronous Overlap-Add). At the final stage, a target F0 contour is predicted either by HMM-based F0 generation or by F0 segment selection; In the case of HMM-based F0 generation, context-sensitive syllable HMMs are used to model and generate expressive F0 contours; In the case of F0 segment selection, an F0 contour is predicted by selecting and concatenating syllable F0 segments from the target emotional corpus using a cost function; the target F0 contour is transplanted onto the second stage synthesized waveform using TD-PSOLA. Their results confirm that F0 segment selection outperforms HMM-based F0 generation. With this conversion approach, they showed that the performance is comparable to a professional voice talent; i.e. the converted neutral utterances were comparable to the same utterances spoken directly in the target emotion.

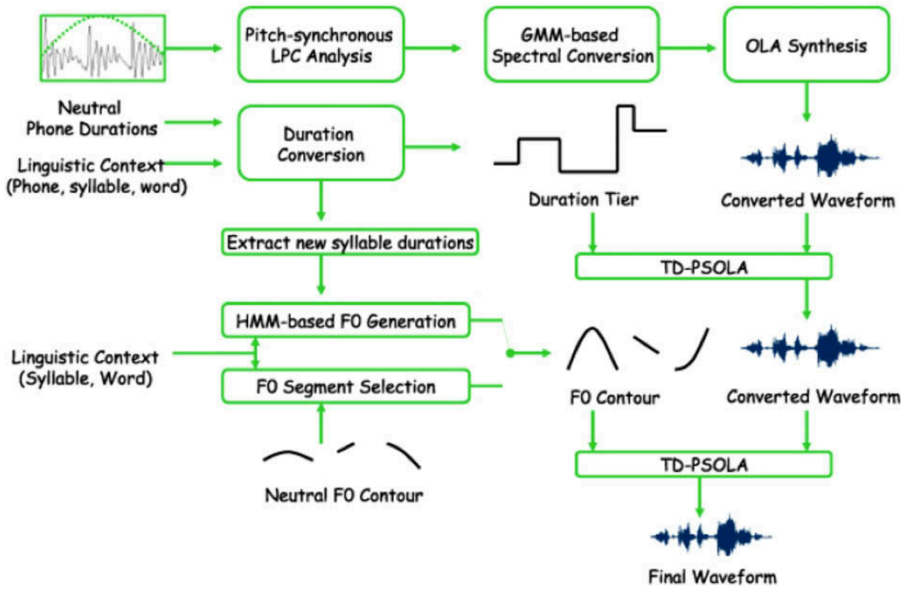


Figure 3.3: An emotion conversion technique proposed by Inanoglu and Young (2009)

3.2.2 Conversational speech synthesis

Campbell (2005) showed that a majority of speech acts in spontaneous speech are vocalizations such as *yeah*, *oh* and *uhuh*. Many of these vocalizations are non-lexical in nature but highly communicative; they are difficult to represent with traditional 'text plus prosody' descriptions. Despite the difficulties involved in data collection of conversational speech, Campbell (2006) emphasizes the importance of conversation-like speech by arguing that phatic communion is as important as propositional content. In his view, attitudes such as *interest* or *boredom* are more relevant to spoken interaction than emotions such as *anger* or *sadness*.

Campbell (2005) claimed that carefully controlled data is not representative of spontaneous speech where many features simultaneously varied to express highly complex communicative intentions. He divided conversational acts, based on whether the utterance is intended primarily to convey information or to display affect, into two types: (i) I-type – to convey or to elicit information; (ii) A-type – to display or elicit display of affect. He proposed a model which can use a huge database of daily interaction with a very complex annotation procedure of A-type utterances. This model gave higher priority to selection constraints instead of 'target cost'; such constraints pri-

marily include, in his words, “CLASS (greet, confirm, complain, filler, laugh, accept, decline etc.,) and VARIANT (happy, sulky, warm, friendly, relaxed, etc.,)" (Campbell 2005, p.380). Then, an optimal token is selected according to continuity constraints. The results has shown unprecedented naturalness of synthetic conversational speech.

Disfluencies (such as *um, uh, ehh, etc.*) are as frequent as the most frequent words in conversational speech, and they carry information and help human communication. Adell (2009) proposed an approach for modeling the prosody of disfluencies. His work analyzes disfluent speech material that includes conversational elements such as filled pauses, repetitions, hesitations and wrappers. He also implemented a multi-layer approach in the Ogmios TTS system (Bonafonte et al. 2006; Bonafonte et al. 2008) in order to synthesize conversational elements. By using context words and part-of-speech (POS) labels of synthetic utterances as features, machine learning techniques such as decision trees and finite state transducers are trained for prosody prediction of disfluencies. This approach showed that the overall naturalness and listening effort of the system is maintained after inserting filled pauses. In other words, his approach increased the spontaneous nature of the speech by keeping the listening effort constant.

Hesitations are another type of disfluencies. Strangert and Carlson (2006) presented an attempt to synthesize hesitation using parametric synthesis. Then, Carlson, Gustafson, and Strangert (2006) investigated features, such as pause duration, retardation and intonation, contribute to the impression of hesitant speech on a surface level. Their perception experiments have indicated pauses and retardations to be among the acoustic correlates of hesitations.

3.3 Interactive speech synthesis: a need for ECAs

Interactiveness has recently become an important requirement for ECAs. It primarily involves *attentive speaking* and *active listening* skills. From the literature, it is evident that very little focus is made on attentive speaking skills of ECAs. In contrast, some significant efforts are made in order to incorporate active listening skills into ECAs. However, incorporation of these skills is an immediate need for ECAs. This section discusses some of these efforts.

3.3.1 Attentive speaking

An attentive speaker pays attention to the listener. He moderates his speech and tailors it to reactions from the listener. In other words, a speaker often will give the listener opportunities for responses, but will also actively receive the responses, and adjust his or her utterances to the occurrence and content of these responses (Bavelas, Coates, and Johnson 2002; Bavelas, Coates, and Johnson 2000). Clark and Krych (2004) identify several ways in which speakers adapt their speech based on opportunities that arise, intentionally or not, mid-sentence. They claim that speakers make the adaptations almost instantly, typically initiating them within half a second of the opportunity arising. The active behavior of the speaker is called as *attentive speaking* (Reidsma et al. 2011).

With the objective of incorporating the attentive speaking skills into Virtual Humans, Reidsma et al. (2011) developed a virtual human platform, Elckerlyc[1] – a new platform for building Virtual Humans. This platform includes a database of motion capture animations containing over 100 direction-giving-task related gestures in the route giving domain. They reported their work that is in progress on several aspects of continuous interaction, such as flexible and adaptive scheduling and planning of multimodal behavior (speech, gestures, facial expressions) including graceful interruption, automatic online classification of listener responses, and models for appropriate reactions to listener responses. As part of their research, several perceptual evaluation studies have been conducted to understand how certain vocal and visual gestures of the agent influence the user's attention in the conversation.

The Listening Talker Project[2] is aimed at many challenges involved in attentive speaking. It is an ongoing EU FP7 project; which has the objective of incorporating attentive speaking skills into speech synthesis systems; which aims to develop the scientific foundations needed to enable the next generation of spoken output technologies; which targets listener-centered speech production systems that are able to work in realistic environments characterized by noise and natural, rapid interactions.

[1] http://elckerlyc.sourceforge.net
[2] http://listening-talker.org

3.3.2 Active listening

The vocal part of the listener's behavior has been extensively discussed in Chapter 2. However, listener responses are usually multimodal in nature. Listeners use audible as well as visible acts to convey their intended meaning. Dittmann and Llewellyn (1968) confirm that most of the listener vocalizations co-occur with visual responses such as head nodes, but interestingly, many research studies on listener behavior were conducted on only one among vocal and visual modalities.

The implementation of the listening behavior has to address several research questions such as *when to trigger a listener response?*, *what to trigger?* and *how to realize it?*. The literature indicates that, nowadays, the focus has been increased in order to investigate these research questions. This section briefly discusses some of the studies.

When to trigger a listener response?

Some research studies showed that there is a strong correlation between triggering of listener behavior and the visible and audible acts (especially, nonverbal behavior) performed by the interlocutor. Ward and Tsukahara (2000) provided evidences for the assumption that the listener's audible acts are often provided when the speaker is able to perceive it more easily. They proposed a model based on acoustic cues to determine the right moment to provide a listener response. The listener's audible acts are provided when the speaker talked with a low pitch lasting 110 ms after 700 ms of speech and provided that such act has not been displayed within the preceding 800 ms. Maatman, Gratch, and Marsella (2005) derived from the literature a list of useful rules to predict when a listener vocalization can occur according to user's acoustic and visual behavior. They also concluded that listener audible and visible acts appear at a pitch variation in the interlocutor's voice.

Bevacqua (2009) implemented a model to trigger visual listener responses according to the user's vocal and visual behavior. She defined a set of probabilistic rules based on the user's multimodal behavior, i.e. visual features such as the head position and orientation; acoustic features such as detection of pause and pitch variation. These rules determine the probability of triggering a listener response.

Morency, Kok, and Gratch (2010) investigated sequential probabilistic models, such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs),

to find opportunities to trigger listener responses. The machine learning models were trained with multimodal features extracted from a database of human-to-human interactions. This approach showed a statistically significant improvement over previously published rule based approaches.

What to trigger?

To determine *which* one of the listener responses the ECA will display, the computational model of Bevacqua (2009) uses the agent's mental state. Her model defines the mental state as a set of communicative functions that the agent wishes to transmit during an interaction. The situational mental state of the agent is formulated by taking into account what the agent thinks and feels about the user's speech. Based on such current state, the agent determines the type of listener response which can communicate the intended intensions. For example, when the agent's mental state is described by the communicative function *agree*, it will display a head nod with an optional smile.

Sevin et al. (2010) described a method to find suitable vocalizations using the user's level of interest and the agent's mental states. The ECA shows its communicative intentions, when it detects that the user loses interest in the interaction, in order to encourage the user to be interested in the interaction. When the ECA estimates that the user's interest level is high and medium/low, it produces mimicry behavior and listener response behavior respectively. If the user's interest level is remarkably low for some time, the agent considers that the interaction is ending and stops progressively doing listener responses.

How to realize listener responses?

The realization of an ECA's listening behavior involves the generation of both vocal and visual modalities. Both modalities usually follow different realization standards.

In Bevacqua (2009)'s work, the realization of visual behavior consists of a behavior planner, a behavior realizer and a FAP-BAP player. With the help of communicative functions involved in the agent's mental state, firstly, the behavior planner plans the listener behavior such as head nod or frown. Secondly, the behavior realizer generates

the animation of the planned behavior following the MPEG-4 format. Finally, the FAP-BAP Player receives the animation generated by the behavior realizer and plays it on a virtual 3D agent.

The literature suggests that the realization of visual behavior is extensively investigated. However, there is not much focus on the synthesis of listener vocalizations. This thesis aims to fill this gap. The investigation for the realizations strategies of the listener's vocal behavior is discussed in next chapters.

3.4 Listening behavior in interactive agents

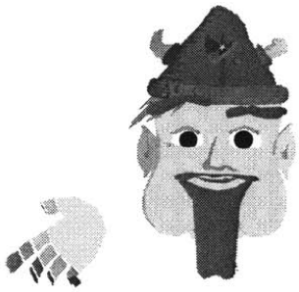
Several interactive agents implemented in the past showed that they have at least minimal listening behavior capabilities. This section discusses such agents.

One of the earliest ECAs was Gandalf (Thórisson 1996), a talking head that has knowledge about the solar system. It has a face and a hand. It was capable of interacting with users using audible and visible acts. It was able to realize several facial expressions and attentional cues. It was able to produce real-time listening behavior based on pause duration information. It was able to display a listener response, a head-nod or a short vocalization, when a pause longer than 110 ms is detected.

Cassell et al. (1999) developed a virtual humanoid, the Real Estate Agent (REA). It shows users the characteristics of houses displayed behind her. She can interact with users through audible and visible behaviors; REA's speech synthesizer allows her to vary the intonation of her voice. Like Gandalf, her listener responses are generated when the user makes a pause longer than 500 ms. Her signals include non-lexical vocalizations such as *mmhmm*, head nods, and short verbal vocalizations such as *I see*.

The Listening Agent (Maatman, Gratch, and Marsella 2005) produces listener vocalizations according to the user's behavior. This agent generates listener responses when it finds lower pitch regions in the speaker's speech. It can produce frowns, body movements and shifts of gaze when the speaker shows uncertainty. It can mimic posture shifts, gaze shift, head movements and facial expressions of the user when listening.

Gratch et al. (2007) developed the "Rapport Agent'", an agent that provides solely visible acts while it is listening. It was implemented to study the level of rapport that users feel while interacting with a virtual dialog partner capable of providing visual



(a) Gandalf

(b) REA

(c) Rapport Agent

(d) MAX

(e) GRETA

Figure 3.4: Multimodal interactive agents (a) Gandalf (Thórisson 1996), (b) REA (Cassell et al. 1999), (c) Rapport Agent (Gratch et al. 2007), (d) MAX (Kopp et al. 2008) and (e) GRETA (Pelachaud 2005)

listener responses. The system uses the user's vocal features for triggering listener responses, whereas it analyses the user's visual behavior, such as head nod, shake, head movement, mimicry, in order to identify what visual signal is to generate. The listener responses comprehend visual signals like head nods, head shakes, head rolls and gaze shifts.

Kopp et al. (2008) developed a virtual human, MAX, in order to make it respond in a pertinent and reasonable way to the statements and the questions asked by a user. The listening model implemented for MAX is based on a reasoning and deliberative processing that plans how and when the agent must react according to its intentions, beliefs and desires. MAX can display multimodal listener responses like head nod,

shake, tilt and protrusion with various repetitions and different movement quality. But they are triggered solely according to the written input that the user types on a keyboard.

A multimodal expressive agent, GRETA[1] (Pelachaud 2005; Niewiadomski et al. 2009) – a real-time three dimensional embodied conversational agent, was developed for conducting research on facial expressions, gestures, gaze, and head movements. Its architecture follows the *whiteboard* design methodology (Thórisson et al. 2005) and is compatible with the standard SAIBA framework (Situation, Agent, Intention, Behavior, Animation) (Vilhjálmsson et al. 2007). To add listening skills to GRETA, Bevacqua (2009) implemented a computational model of listening behavior in its framework as described above. This model automatically computes *when* a backchannel signal should be emitted; and *which* communicative function the agent should transmit through its listener behavior. The GRETA agent can realize around twelve communicative functions through its visual listener behavior: they are *agree, disagree, accept, refuse, believe, disbelieve, interest, not interest, like, dislike, understand* and *not understand*.

3.5 Summary

This chapter covered the different notions of the current day speech synthesis technologies; It also focused on the vocal listener behavior of interactive agents. We briefly described unit-selection and HMM-based speech synthesis technologies. We reviewed several interesting research studies on emotional and conversation-like speech synthesis. We explained the need for attentive speaking and active listening skills in order to improve interactiveness of ECAs. Finally, we also reviewed several interactive agents that are able to generate and realize the listener's behavior.

[1] http://perso.telecom-paristech.fr/ pelachau/Greta

Chapter 4

Frameworks

The major objectives of this thesis, as described in Chapter 1, are: (i) endowing TTS with the capability to synthesize listener vocalizations; (ii) integration of vocal listener behavior into the SEMAINE framework in order to make Sensitive Artificial Listener (SAL) characters 'listen' actively. To achieve them, this thesis uses two open source frameworks. One is a text-to-speech framework – The MARY TTS platform[1]; the other one is a multi-modal interaction framework – The SEMAINE framework[2]. Both of them were designed to promote ongoing research in their respective domains. This chapter discusses the two open source frameworks.

The chapter is primarily divided into two sections. Section 4.1 focuses on the MARY TTS platform, particularly: (a) its architecture; (b) the procedure to build voices; (c) generation of synthetic speech at runtime. Section 4.2 discusses the SEMAINE framework aimed to demonstrate SAL characters, but the primary focus is on the listener behavior of these characters. In addition, we briefly describe SEMAINE API that is used in this framework. Finally, this chapter is summarized in Section 4.3.

4.1 The open source Mary TTS platform

The current architecture of the open source MARY (Modular Architecture for Research on speech sYnthesis) platform is shown in Figure 4.1. MARY is a stable Java server capable of multi-threaded handling of multiple client requests in parallel. The design is

[1] http://mary.opendfki.de
[2] http://semaine.opendfki.de

highly modular: a set of configuration files, read at system startup, define the processing components to use. For example, the file `de.config` defines the German processing modules, while `en_US.config` defines the (US) English modules. If both files are present in the configuration directory, both subsystems are loaded when starting the server. Each synthesis voice is defined by a configuration file: `de-bits1.config` loads the unit selection voice bits1, `de-bits1-hsmm.config` loads the HMM-based voice bits1-hsmm, etc. The MARY framework allows a step-by-step processing with an access to partial processing results. This framework is composed of distinct modules and has the capability of parsing speech synthesis markup such as SSML[1] (Speech Synthesis Markup Language). More details on the MARY architecture can be found in Schröder and Hunecke (2007).

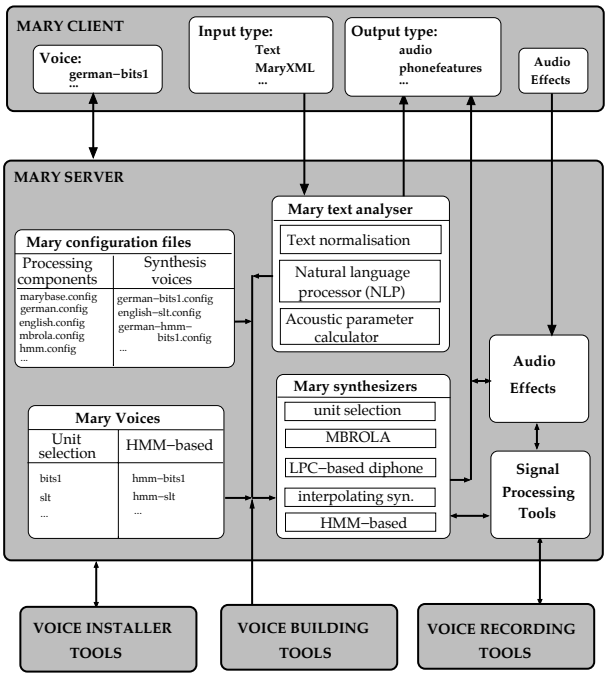


Figure 4.1: Mary TTS platform version 4.0

Currently, the list of available waveform synthesizers includes a unit selection synthesizer (Schröder, Hunecke, and Krstulovic 2006), an MBROLA diphone synthesizer (Dutoit et al. 1996), an experimental interpolating synthesizer (Schröder 2007) and a

[1] http://www.w3.org/TR/speech-synthesis11

new HMM-based synthesizer ported to Java from the excellent HMM-based synthesis code from the HTS project[1] (Tokuda et al. 2008). The MARY text analyzer components are described in (Schröder and Trouvain 2003). The audio effects component is a new component designed to apply different effects on the audio produced by the different synthesizers. The effects are set through the audio effects GUI of the MARY client component. The Voice installer tools component is used for downloading and installing new voices or removing already installed ones. The voice recording tool is a component designed to facilitate the creation of speech synthesis databases. The voice building tools are a set components used to build new voices.

The workflow of MARY framework can be divided into two stages: (i) voice building process; (ii) runtime synthesis. The following sections explain these stages in detail.

4.1.1 Voice-building process

The steps required to build a new voice of a new language from scratch are illustrated in Figure 4.2. Two main tasks can be distinguished: (i) building at least a basic set of natural language processing (NLP) components for the new language, carrying out tasks such as tokenization and phonemic transcription (left branch in Figure 4.2). and (ii) the creation of a voice in the language (right branch in Figure 4.2).

Whereas high-quality support of a language will usually require language-specific processing components, it is often possible to reach at least a basic support for a language using generic methods (Black and Lenzo 2003). Once the NLP components have been developed, the task of creating a voice can be pursued (right branch in Figure 4.2). First, a recording script providing good diphone and prosodic coverage is selected from the text collection. Using the NLP components a *feature maker* component annotates each sentence in the text database with diphone and prosody features to be used in a greedy selection. The resulting collection of sentences can be used as the recording script for voice recordings with the tool *Redstart*. The recorded audio files can then be processed by the MARY voice import tools which generate a unit selection and/or an HMM-based voice, as well as speaker-specific prediction components for acoustic parameters. If, during the voice-building process, force-aligned transcriptions

[1] http://hts.sp.nitech.ac.jp

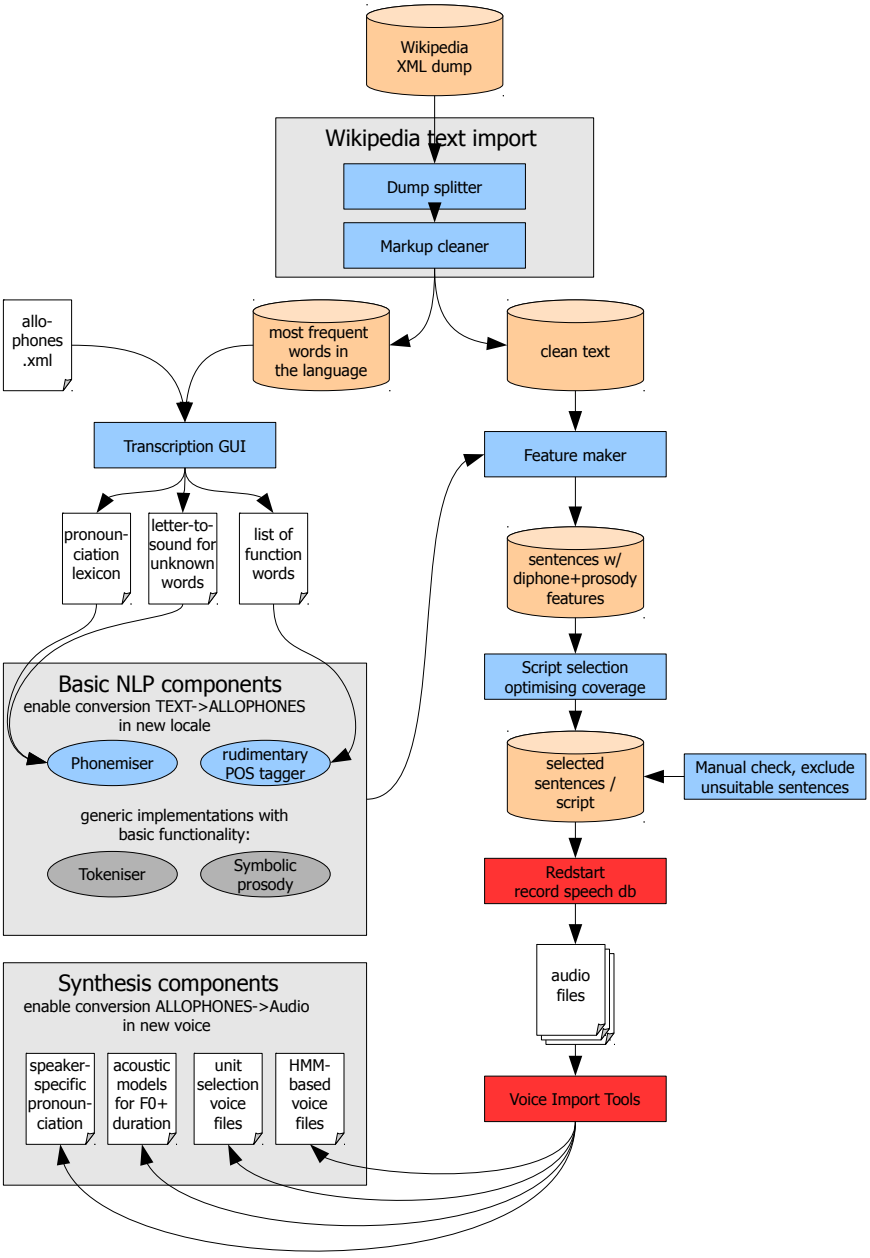


Figure 4.2: Workflow for multilingual voice creation in MARY TTS (Pammi, Charfuelan, and Schröder 2010), more information about this tool can be found in: *http://mary.opendfki.de/wiki/NewLanguageSupport*

were manually corrected, it is also possible to predict *speaker-specific pronunciations*. In the following these steps are explained in more detail.

Preparing NLP components for a new language

The preparation of NLP components is required only if the language is unknown (new) to the MARY server. To support a new language, the workflow starts with a substantial body of Unicode text in the target language, such as a dump of Wikipedia in that language. After automatically extracting text without markup, as well as the most frequent words, the first step is to build up a pronunciation lexicon, letter-to-sound rules for unknown words and a list of function words. The framework provides a *Transcription GUI*, which a language expert can use to generate a pronunciation dictionary. An `allophones.xml` file defines the allophones of the target language that can be used for transcription, and it characterizes them using a set of phonetic features. The features include length, height, frontness and lip rounding for vowels, as well as type, place of articulation and voicing for consonants. If other features are distinctive in a given language, additional features can be added without any problems. The allophones file has to be prepared manually by a language expert.

Once the allophones file for a target language is available, a language expert or at least a native speaker can use the Transcription GUI to transcribe as many of the most frequent words as possible using the allophone inventory. The tool supports this task by training, on the available data, a letter-to-sound predictor which can propose candidate transcriptions for untranscribed words. Furthermore, it is possible to mark function words in the list in order to enable a simplistic POS tagger, which works based on simple context-free string matching. Where a better quality POS tagger or morphological analysis is required, a custom TTS module needs to be implemented. This is unproblematic due to the modular architecture of the MARY TTS system.

With this minimal manual input for a new language, a simple NLP system can be built, using a generic tokenizer and a rule-based prediction of symbolic prosody.

Optimal text selection

Creating a recording script that provides a good diphone and prosodic coverage is not a trivial task. In the MARY voice creation toolkit a greedy algorithm is used for

selecting sentences to optimize coverage. Three parameters are taken into account: the units, coverage definition and stop criteria. Units are defined as vectors consisting of three features: phone, next phone and prosody property. The definition of coverage fixes which kind of units are wanted in the final set; in the current version all diphones and their prosodic variation are used. Other aspects like frequency weights, sentence length, features weight, etc. can be set for optimizing the coverage. The stop criteria are a combination of number of sentences, maximum diphone coverage and maximum prosody coverage (Hunecke 2007). The selected sentences then need to be manually checked in order to discard any problematic sentences – e.g., sentences that are too long or that contain words that might be too difficult to pronounce fluently.

If the aim is to support a specific domain, it is possible to either use domain-specific material instead of general-domain text as the basis for selection, or, if the domain is small enough, to manually design a representative set of sentences.

Voice recording

MARY comes with a tool called Redstart to assist the user in the process of voice recording. The tool displays sentences one by one, and records each sentence into a separate wave file. An estimate of recording time is used to pace the recordings; beep sounds indicate when the microphone is opened and closed. Checks for temporal and amplitude clipping are automatically performed; if in doubt regarding the quality of a recording, the user can play the recorded waveform, display the speech signal and the corresponding spectrogram, pitch, and energy contours, and of course re-record the sentence. No files are overwritten, a history of attempts to utter a given sentence is kept. This way, it is possible to revert to the best recording achieved rather than having to try until a perfect version is produced.

Voice import components

The voice import tool combines an extensible list of components in a simple GUI[1] (see Figure 4.3), designed primarily to facilitate the creation of new voices by users without expert knowledge of speech synthesis. The user can select a series of import components, which are run in sequence. A progress bar is shown for the component which

[1] http://mary.opendfki.de/wiki/VoiceImportComponents

is currently running. After successful completion, the component is colored in green; if processing fails, it is displayed in red, and processing of subsequent components is aborted.

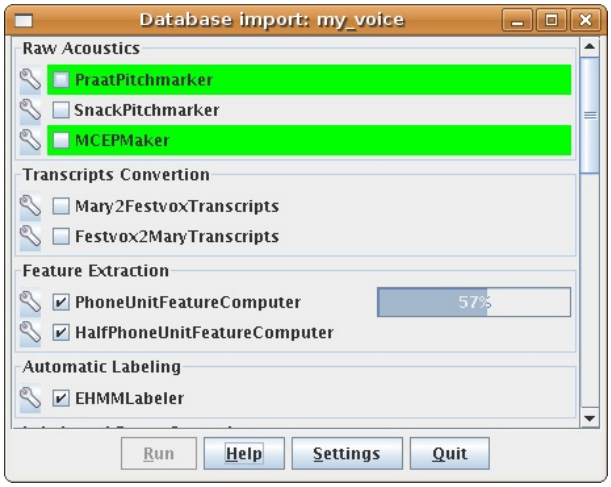


Figure 4.3: MARY voice import tool components (Pammi, Charfuelan, and Schröder 2010)

Several voice import components execute high quality, freely available components specialized for particular tasks, for example, for automatic labeling we can use Festvox's[1] EHMM (Anumanchipalli, Prahallad, and Black 2011), or for training HMM models we use the scripts provided by HTS adapted to the MARY TTS architecture. This toolkit provides reasonable baseline configuration settings to external tools to allow non-expert users to execute the tools in a default setting; experts are given the option to configure many aspects if needed.

Quality control on automatic labeling

In unit selection and HMM-based speech synthesis systems, accurate phonetic segmentation (labeling) is required to ensure quality of speech. The quality of labeling determines the quality of units, which might be affected by a range of problems including misaligned phone boundaries, mismatches between the phones that are labelled and that are pronounced, and the presence of background noise. Estimating the

[1] http://festvox.org/

quality of individual units in the database is a key issue in order to reduce the amount of manual correction effort or as a criteria to apply when choosing a unit during synthesis. This toolkit provides a component to estimate the quality of labeling using a statistical model cost measure, comparing recorded phones to "average" acoustics as generated by an HMM synthesis model trained on the same data (Pammi, Charfuelan, and Schröder 2009). This component estimates quality of individual phonetic segments. When a human inspects the labels in the order given by this component, more errors can be found in a given time than with simple linear inspection.

Creation of unit-selection voices

In MARY framework, a set of voice import components is dedicated to create unit selection voices. The objective of these components is to create a diphone unit database that is usable for the runtime synthesis framework. Firstly, the components compute pitchmarks and pitch-synchronous Mel Frequency Cepstral Coefficients (MFCC) vectors with Praat[1] (Boersma and Weenink 2010) and the Edinburgh Speech Tools[2] (EST) (Taylor et al. 1999), respectively. Secondly, they predict linguistic feature vectors with the MARY system; acoustic unit features (F0 and duration) are added based on pitchmarks and on automatically labelled phoneme boundaries. Then, they measure join cost features (Fundamental Frequency (F0) and MFCCs) at the first and last frame of each halfphone unit. In the later stages, the components construct Classification and Regression Trees (CART) to predict unit duration and F0 using the EST tool *wagon*. Finally, *CARTBuilder* component builds a pre-selection tree in a two step procedure: (i) build a hard coded 'top-level' tree with phonetic properties such as phoneme identity, stress status, voicing etc. (ii) construct bottom-level tree by an automatic tree-growing procedure for each of the top-level leaves, using acoustic distance between units as the impurity measure; the acoustic distance between two units is measured as the weighted sum of differences in duration, F0, and spectral difference computed as the Mahalanobis distance of MFCC vectors.

Once the voice building process is completed, the voice installer component can install the unit-selection voice into MARY server. This voice includes the pre-selection

[1] http://www.fon.hum.uva.nl/praat/

[2] http://www.cstr.ed.ac.uk/projects/speech_tools/

tree, F0 and duration CART trees, an audio timeline file that contains labelled speech segments, and the unit and feature files needed to compute target and join costs.

Creation of HMM-based voices

For creating HMM-based voices, voice import components use a version of the speaker dependent (or adaptive) training scripts provided by HTS (Tokuda et al. 2010), adapted to the MARY platform. The HMM training scripts and programs have been slightly modified from the original HTS scripts, basically they have been changed to use context features predicted by the MARY text analyzer instead of the Festival Speech Synthesis System (Black, Taylor, and Caley 1998).

The MARY 4.0 release provides a patch file to be applied to the HTS training scripts so the MARY voice building tools can be used in a straightforward way. The main changes included in this patch are:

- Calculation of bandpass voicing strengths for mixed excitation (Yoshimura et al. 2001a); modifications in training scripts to consider them in order to make sure that the generated parameters include bandpass voicing strengths at runtime.
- Extraction of monophone and fullcontext labels from MARY context features.

The current procedure for creating a new HMM-based voice can be summarized in three steps: data preparation, training of HMM models and installation of a new voice into the MARY system. MARY uses the HTS standard training procedure, where the spectrum is modeled by generalized mel-cepstral coefficients, the excitation part is modeled by log fundamental frequency (log F0) and state durations of each HMM are modeled by a multivariate Gaussian distribution (Yoshimura et al. 1998). In addition, the procedure includes bandpass voicing strengths parameters for modeling mixed excitation as reported in Yoshimura et al. (2001a). Five bandpass filters are used to generate these strengths.

4.1.2 Runtime synthesis in MARY platform

In the previous section, we have provided information on MARY-based off-line processing methods for voice preparation. This section discusses runtime synthesis methods used in MARY framework for unit-selection and HMM-based voices.

Unit selection synthesis

The unit selection system in MARY implements a generic unit selection algorithm, combining the usual steps of tree-based pre-selection of candidate units, a dynamic programming phase combining weighted join costs and target costs, and a concatenation phase joining the selected units into an output audio stream. The unit selection framework uses diphone units during concatenation phase, because joining in the mid-section of phonemes is expected to introduce less discontinuities than joining at phoneme boundaries. For each target diphone, a set of candidate units is selected by separately retrieving candidates for each halfphone through a decision tree, and retaining only those that are part of the required diphone. When no suitable diphone can be found, the system falls back to halfphone units.

The most suitable candidate chain is obtained through dynamic programming, minimizing a weighted sum of target costs and join costs. Both are themselves a weighted sum of component costs. On the one hand, target costs cover the linguistic properties of units, and the way they match the linguistically defined target. It also includes acoustic target costs used for comparing a unit's duration and F0 to the ones predicted for the target utterance by means of regression trees trained on the voice data. On the other hand, join costs are computed as a weighted sum of F0 difference and of spectral distance, computed as the absolute distance in 12-dimensional MFCC space.

The challenge in all unit selection systems is determining appropriate weights for the individual target and join cost components. However, MARY does not have a principled way of determining these weights. Therefore, we have to set a number of ad hoc values through iterative listening and adapting. After the chain of units minimizing these costs is determined, the units are retrieved from a timeline file and concatenated using overlap-add of one pitch period at the unit boundaries.

HMM-based synthesis

During HMM-based synthesis the text analyzer of the MARY server converts the text into a context-based label sequence, HTSCONTEXT format in MARY platform. This sequence is passed to a HMM-based synthesizer which is based on a Java port of the hts_engine (Tokuda et al. 2010). The context-based label sequence is converted into a sequence of context dependent HMMs. State durations of each model in this sequence

are estimated from the Gaussian distributions. The next step is to generate the parameters: mel-cepstral coefficients, log F0 values and bandpass voicing strengths using the maximum likelihood parameter generation algorithm including global variance (Toda and Tokuda 2005). Once the parameters have been generated, F0 parameters, bandpass voicing strengths and the five passband filters (the same used for generation of bandpass voicing strengths, during training) are used to generate mixed excitation as in (Yoshimura et al. 2001a). Finally, speech is synthesized from the mel-cepstral coefficients and mixed excitation values by using the MLSA filter (Tokuda, Zen, and Black 2002).

4.2 The SEMAINE framework

The main objective of the SEMAINE framework is to build a Sensitive Listener Agent (SAL) – a virtual agent who pretends to be an emotional being; who interacts with the help of analyzing the user's nonverbal behavior in his/her visible and audible acts. This framework consists of a SAL architecture that is built on top of the SEMAINE API (Schröder 2010), an open source framework for building emotion-oriented systems. This section provides background information on two major parts of SEMAINE framework: the SEMAINE API and the SAL architecture.

4.2.1 The SEMAINE API

The SEMAINE API is an distributed multi-platform component integration framework for real-time interactive systems, see Figure 4.4.

As shown in Figure 4.4, the architecture of SEMAINE API uses a message-oriented middleware (MoM) (Banavar et al. 1999) in order to integrate several *components* – where actual processing of the system is defined. Such *components* communicate via a set of *topics* (i.e. `semaine.data.*`). Here, a *topic* is a virtual channel where each and every published message, addressed to that *topic*, is delivered to its subscribed consumers. Each of the components sends its meta information to the system manager via its meta messenger and the topic `semaine.meta`. Each component can optionally publish log messages to a set of topics `semaine.log.*`; whereas, a configurable

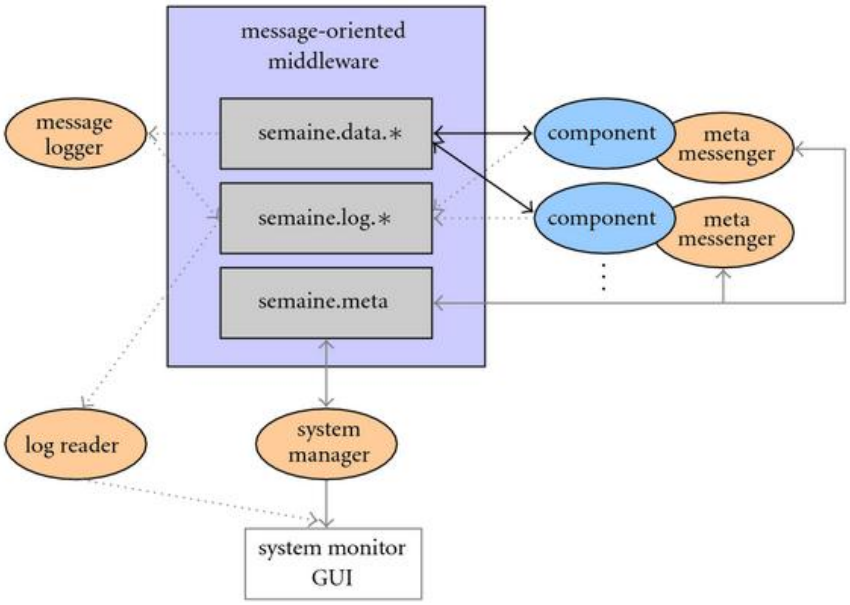


Figure 4.4: The SEMAINE API architecture (Courtesy from Schröder 2010)

log reader can receive the log messages from those topics. A system monitor GUI displays, on the one hand, a flow graph of components based on the information received from the system manager. On the other hand, the GUI shows log messages obtained from the log reader – such mechanism is beneficial for debugging the real time system.

Technical details of the API

The communication passes via the message-oriented middleware ActiveMQ[1], which is reasonably fast and supports multiple operating systems and programming languages. For component integration, the SEMAINE API encapsulates the communication layer in terms of components that receive and send messages, and a system manager that verifies the overall system state, provides a centralized clock independent of the individual system clocks. The API makes it particularly easy for components to communicate via a number of standard representation formats such as the Behavior Markup Language (BML)(Vilhjálmsson et al. 2007), but also allows for arbitrary messages so

[1] http://activemq.apache.org

that the functionality can be easily extended. The platform is publicly available as open source; detailed information about its extensibility is available (Schröder 2010).

4.2.2 The architecture of the SAL system

The conceptual architecture of the SAL system is shown in Figure 4.5. Components are shown as ovals, message types as white rectangles. The raw user input is converted by a set of feature extractors into raw feature vectors which are sent very frequently (e.g., every 10 ms for audio, and for every video frame). The 'Analysers' are components such as classifiers which derive some sense from the raw features in a context free manner; 'Interpreters' are then considering the analysis results in the light of everything the agent knows about the current and recent state of the world, and ultimately derive the system's "current best guess" regarding the state of the user and the dialogue, and update the agent's own state in the light of this evidence. In parallel, the 'Intent Planner' can continuously propose appropriate actions, which are filtered by an 'Action Selection' before they are realized as agent behavior. In order to realize an action, the multimodal behavior is planned based on a representation of the communicative function, realized in terms of the synthetic audio and synchronized player directives for the visual behavior, and finally it is rendered by a 3D character player.

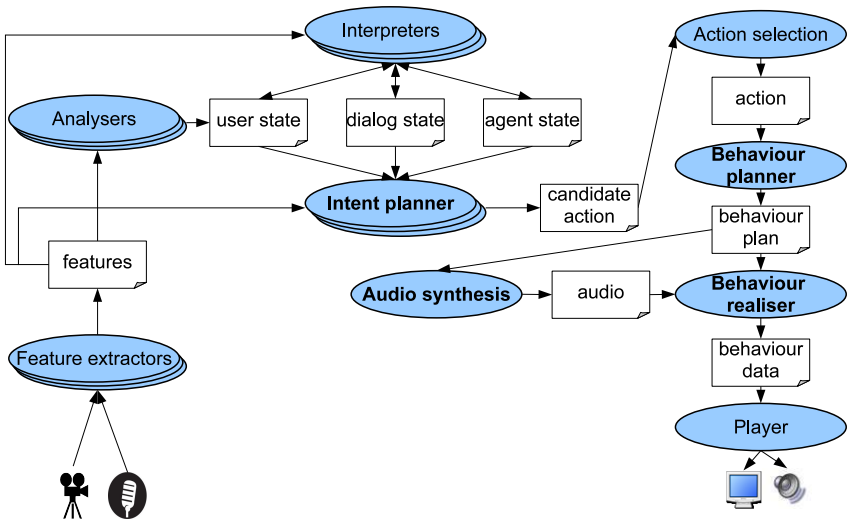


Figure 4.5: Architecture of the SAL system

The architecture can generate the agent's behavior both while it speaks and it listens. This thesis focuses on listener behavior, so we present in more details how the modules of the architecture generate this type of behavior. The concerned modules, namely the Intent Planner, the Audio Synthesizer, the Behavior Planner and the Behavior Realizer, are highlighted in bold in Figure 4.5.

Listener Intent Planner

The Intent Planner is composed by two fundamental sub-modules: the speaker and the listener intent planners. The Listener Intent Planner computes the agent's behavior while being a listener when conversing with a user.

Chapter 3 discussed that there is a strong correlation between listener responses and the acoustic and visual behaviors performed by the speaker. This component contains a set of probabilistic rules to decide *when* a listener response should be triggered; these rules are formulated from previous research studies (Maatman, Gratch, and Marsella 2005; Ward and Tsukahara 2000).

The SAL system analyzes user's behaviors looking for those that could prompt an agent's signal; for example, a head nod or a variation in the pitch of the user's voice will trigger a listener response with a certain probability. Then, the system calculates *which* listener response should be displayed. The agent can provide either *response* signals that transmit information about its communicative functions (like agreement, liking, believing, being interested and so on) (Allwood, Nivre, and Ahlsén 1992; Poggi et al. 2005) or signals of mimicry that mirror the speaker's signals. The Action Selection module receives all the candidate actions and selects which one will be actually displayed according to the user's interest level (Sevin et al. 2010).

Behavior Planner

This module receives as input the agent's communicative functions specified in the listener's *response* and some agent's behavioral characteristics (i.e. *baseline*). Its task consists in generating a list of adequate behavioral signals for each communicative function. The agent's *baseline* contains information on the preference the agent has in using a modality (speech, head, gaze, face, gesture, and torso) (Mancini et al. 2008). Apart from the speech modality, the baseline specifies also the expressive quality for

each other modality. Expressivity is defined by a set of parameters that affect the qualities of the agent's behavior: e.g. wide vs. narrow gestures, fast vs. slow movements. All the possible listener's communicative functions are associated with the multimodal signals that can be produced by the listener in order to convey them. Each of these associations represents one entry of a lexicon, called *backchannel lexicon*. Depending on the agent's baseline and the communicative function to convey, the system selects in the lexicon the most appropriate multimodal behavioral set to display. For example, an agent that wants to communicate its agreement could simply nod, or nod its head and smile, or say *m-mh*.

Audio Synthesis

The implementation of this module is one of the major objectives of this thesis. The responsibility of this module is not only synthesis of spoken utterances, but also synthesis of vocal listener behavior which includes listener vocalizations like *myeah*, *uh-huh* and *oh* etc. For better lip synchronization of audiovisual vocalizations, this module uses a speech synthesis system, which will be implemented in MARY framework, for generating the speech with timing information. This module receives as input the agent's planned behavioral characteristics in a form of markup request; then, the synthesis system looks up available vocalizations for the given speaker and generates the most appropriate vocalization found for the request.

Behavior Realizer

This module generates the animation of our agent following the MPEG-4 format (Ostermann 2002). The input of the module contains the verbal and visual signals selected by the Behavior Planner. Facial expressions, gaze, gestures, torso movements are described symbolically in repository files. Temporal information about the vocalization generated by the Audio Synthesis module, are used to compute and synchronize lips movements.

4.3 Summary

This chapter outlined two frameworks which will be used in this thesis. They are the MARY TTS platform and the SEMAINE framework. We started with MARY speech synthesis framework: its architecture, its voice building workflow, and its runtime synthesis approaches such as unit selection and HMM-based synthesis techniques. Then, we briefly described a multicomponent integration platform for realtime interaction systems – the SEMAINE API – used to build SAL agents. We finally discussed the SAL architecture which includes several multimodal analysis and synthesis components; however, we mainly focussed on realization of the listener's behavior.

Part II

Investigation

Chapter 5

Methodology

The previous chapters have explained the relevant background literature required for the thesis. The rest of the chapters describe an investigation for generating listener vocalizations. This thesis is the first attempt to incorporate the ability to synthesize natural listener vocalizations in a full-scale speech synthesis system. Therefore, a systematic methodology is needed for the investigation. This chapter discusses our methodology for the investigation.

In Section 5.1, we start with identifying challenges involved in corpus-driven speech synthesis techniques to synthesize listener vocalizations. Section 5.2 list out research questions needed to address in order to achieve the objectives of this research. Section 5.3 proposes a methodology to find solutions to the research questions.

5.1 Challenges involved in synthesizing vocalizations

In this section, we discuss major challenges in corpus-based techniques to synthesize listener vocalizations: *how to record high quality natural listener vocalizations? what are the best symbols to represent them? what is the best algorithm to realize appropriate listener vocalizations?*

5.1.1 Corpus collection

Corpus-based unit selection TTS systems reach highly natural expressivity, by recording a separate voice database with the same speaker for each targeted expressive tone.

Therefore, the naturalness of the synthetic speech depends on the quality of recordings. The set of sentences used for such recordings is pre-defined, and selected from a large text corpus with optimal coverage of phones.

In the case of listener vocalizations, many of them are nonlinguistic in nature like *laughter*, *sighs* and *hmm*. It is very clear that we cannot use the usual criterion of optimal phone coverage to generate these vocalizations. Moreover, pre-defined scripts will be unsuitable to collect natural listener vocalizations, because listener vocalizations are likely to occur naturally only in conversation.

The information of listener vocalizations is stored in several behavioral properties like segmental form, intonation, and voice quality. These vocalizations are context dependent in dialogue; and up to some extent they are speaker dependent as well. Therefore, recording vocalizations with all possible combinations of behaviors with a single speaker is almost impossible. In other words, the acoustic variability of recorded listener vocalizations is expected to be limited.

5.1.2 Symbolic representation

As described in Chapter 4, traditional speech synthesis systems represent speech material as a sequence of phonetic symbols. An automatic force-alignment technique can be used to time-align phonetic labels with the audio signal, called *phonetic labelling*. In the case of listener vocalizations, the meaning of vocalizations depend not only on the phonetic sequence (segmental form) of vocalizations, but also on several prosodic properties like intonation and voice quality. Moreover, getting phonetic representation and its timing information of non-linguistic vocalizations, like *laughter* and *sighs*, are very difficult due to their non-lexical nature.

According to Chapter 2, the *meaning* conveyed by a vocalization is an important property that needs to be annotated. However, no standard approach for meaning annotation of vocalizations is available in the literature. In addition, we do not know the suitable meaning descriptors to represent all listener vocalizations.

5.1.3 Realization algorithm

The synthesis algorithm is expected to handle the situation of limited acoustic variability. The challenge for such algorithm is not only synthesis of high quality natural

listener vocalizations, but also realization of appropriate vocalizations for given user requests. Such requests contain the intended meaning of the user. For a given limited corpus of listener vocalizations, we have to investigate strategies to synthesize an appropriate vocalization for a given user request without dropping the naturalness of originally recorded vocalization.

5.2 Research questions

To overcome the above mentioned limitations for synthesizing listener vocalizations, we start by phrasing the challenges involved in terms of research questions. The main research questions to address limitations for generating listener vocalizations are the following:

- How to collect a database of listener vocalizations?
- What kinds of meanings are expressed through listener vocalizations?
- What form is suitable for a given meaning?
- How to annotate meaning and behavior (form) of a listener vocalization?
- How to realize the form using a technological framework?

As the synthesis of listener vocalizations is a new topic in synthesis, we are not aware of any technological framework to synthesize these vocalizations. To come up with a new framework to realize vocalizations, some technological research questions should be answered like:

- What kind of technology is suitable to synthesize nonverbal vocalizations? Unit selection, HMM-based or other?
- If it is unit selection, what strategy would be suitable to select unit?
- If it is HMM-based, how to model and realize nonverbal vocalizations?
- How to get advantage from signal modification algorithms?

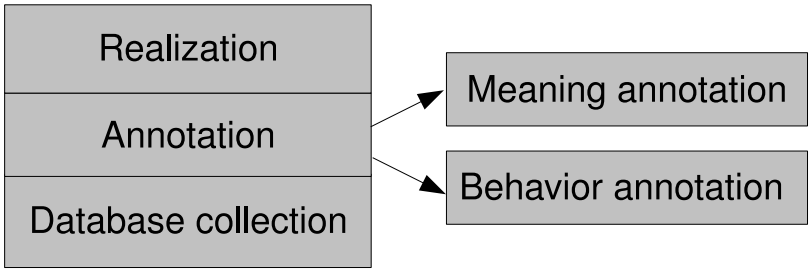


Figure 5.1: Major aspects of proposed work

The major objective of this work is not only providing answers to the above research questions, but also building a system to synthesize nonverbal listener vocalizations and adding a new functionality to text-to-speech synthesis that can synthesize nonverbal vocalizations. The system has to be robust and real-time compatible with a multi-modal synthesis system and it has to use standard representations like eXtensible Markup Language (XML) formats in view of future inter-module communication. A possibility is there to raise more research questions when we try to evaluate the final system on a real-time multi modal interaction system.

5.3 Proposed methodology

We have described the SEMAINE framework in Chapter 4. In the framework, *Listener Intent Planner* plans not only the intentions of the SAL agent while listening, but also the timing information to trigger their behavior. The description of listener intention uses standard XML representation ('Multi-modal XML input' in Figure 5.2). The objective of this thesis is to synthesize appropriate listener vocalizations for such markup requests.

This section describes the conceptual model of our proposed methodology to build a framework for synthesis of listener vocalizations. The proposed work consists of three different levels (as shown in Figure 5.1): Data collection, Annotation and Realization.

5.3.1 Database collection

As the traditional recording setup is not useful to capture listener vocalizations, we propose to record natural dialog speech between an actor and his/her dialog partner in an anechoic studio because listener vocalizations seem to be natural only in a conversation. According to the new proposed recording setup, the actor and the dialog partner will sit in different rooms and hear each other using headphones, so that we can record each speaker's voice on a different channel without interference of the other speaker's speech. As we are aiming to capture listener vocalizations, the actor will be instructed to participate in a free dialog, but to take predominantly a listener role.

5.3.2 Annotation

In order to determine different kinds of meanings expressed through listener vocalizations, the perceived meaning behind each vocalization should be annotated. Similarly, the annotation of behavioral properties will be useful to know identify behavior for a given meaning. Initially, we do not know how many meaning or behavior categories are necessary to annotate all listener vocalizations, so we propose to annotate all vocalizations using informal descriptions to make sure that we are not guided by any pre-existing set of categories. These categories may or may not be suitable to represent all listener vocalizations available in our data. So informal descriptions will be helpful to understand better the structure of both behavior and meaning. Subsequent grouping of these descriptions will help to understand the types of behavior and meaning of listener vocalizations, at least for the speaker we studied. In the later stages, a suitable limited set of categories that capture the essence of meaning as recorded in informal descriptions will be identified.

The sequence of steps involved in the proposed annotation scheme is the following. Firstly, start-end time labels will be annotated for all listener vocalizations made by the actor. Secondly, informal descriptions will be provided for each labeled segment in three different levels: content, behavior, sub-texts. In later stages, suitable meaning category will be identified for each vocalization with the help of informal descriptions. Finally, annotation for behavioral properties like intonation, voice quality etc. will be provided.

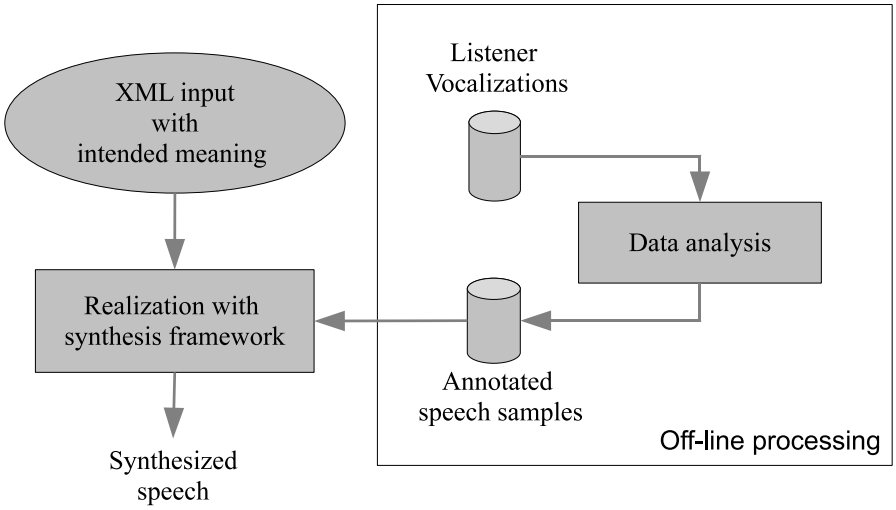


Figure 5.2: Realization methodology

5.3.3 Realization

The conceptual model for the realization system, as shown in Figure 5.2, contains off-line and runtime processing modules. Data analysis on annotated speech samples is a crucial step in off-line processing which provides relations between behavior and meaning. The experience from this analysis will let us know whether the relation between meaning and behavior is a one-to-one mapping or whether a single behavior can be usable to simulate multiple intended meanings. Substantial work is expected on the level of the technological framework to realize a nonverbal listener vocalization. For example, we have to find a way to model and generate nonverbal vocalizations if we choose Hidden Markov Model (HMM) based synthesis as a technological framework.

The proposed runtime system will work as follows. Initially, an XML front-end processing module will identify the intended meaning behind requested nonverbal vocalizations. The next module will be finding suitable behavior to the requested meaning category with the knowledge of relations between behavior and meaning. Finally, another module will realize the appropriate behavior with a synthesis technology like unit-selection or HMM-based synthesis.

5.4 Summary

This chapter discussed the major limitations for synthesizing listener vocalizations. By identifying a number of relevant research questions to investigate, it proposed a three-stage methodology: (i) *data collection*, (ii) *annotation*, and (iii) *realization*. The solutions identified from the proposed research work will lead us towards expressive conversational speech synthesis. The main contribution of this research work is not only providing technological solutions to generate nonverbal listener vocalizations, but also building a real-time system that can be integrated with the SEMAINE project demonstration system which is aiming to build an audiovisual SAL system.

Chapter 6

Database collection

Nowadays high quality speech synthesis systems use data-driven approaches. Unit-selection and HMM-based synthesis technologies are well known examples. In either of these technologies, the corpus plays a key role in the quality of TTS systems. In order to qualify for building a high quality voice, the corpus has to satisfy basic requirements. In unit-selection speech synthesis, for example, a fundamental requirement is that the corpus should be recorded in an anechoic chamber for noise-less speech.

The aim of this chapter is to explore research questions involved in collecting listener vocalizations and discuss possible solutions. With the known importance of basic requirements, this chapter starts with collecting key requirements to be satisfied (in Section 6.1). Section 6.2 describes the need to collect a new database. In Section 6.3, we proposed a method to acquire listener vocalizations. Section 6.4 discusses the first experimental collection of German listener vocalizations from a professional German actor. In Section 6.5, we explain our efforts to collect British English vocalizations from four professional British actors.

6.1 Requirements

Data requirements are crucial to achieve high quality in data-driven approaches. In this research, the requirements considered not only objectives of speech synthesis specific research, but also goals of the SEMAINE project for the best possible performance of its demonstration system. This Section list all possible requirements for corpus collec-

tion. While synthesis specific requirements aim for better quality interactive synthesis system, the project specific requirements target interfacing challenges of project's feedback mechanism.

6.1.1 Generic requirements

- ***High quality*** As the speech material used for synthesis should be high quality, we have to record vocalizations in an anechoic chamber.

- ***Naturalness*** The aim of this research is to support better interactive speech synthesis. Such support is possible only when listener vocalizations are natural enough for interactive environments.

- ***Acoustic variability*** In order to synthesize vocalizations with many possible meanings, we must first of all record a wide variety of vocalizations with the maximum amount of acoustic variability.

6.1.2 Project specific requirements

- ***Record from whom?*** The corpus of listener vocalizations have to be recorded from the same speaker with whom a speech synthesis database is recorded to create the voice of a given SAL character, so that it is possible to generate both speaking and listening behavior with the same voice.

- ***Chatting Scenario*** The development of the Sensitive Artificial Listener (SAL) scenario is based on the idea of a chat system that engages users by encouraging them to talk more, using stock phrases and follow-up questions. The listener vocalizations that are recorded in data collection should be natural enough to be used in this scenario.

- ***Four SAL Characters*** The Sensitive Artificial Listener (SAL) system that aims to create four different emotional characters demands listener vocalizations with four different styles: pragmatic, happy, gloomy or aggressive styles. We have to record data from either four different speakers or a single speaker who can produce vocalizations with different speaking styles.

6.2 Need for a new corpus

Although several databases that contain natural listener responses are publicly available, they are unable to fulfill the basic requirements of speech synthesis databases. The AMI meeting corpus (Carletta et al. 2006), for example, is a multi-speaker meeting corpus, where the quality of speech material is not up to the mark of speech synthesis requirements. The available listener responses in this corpus are unsuitable for voices developed for SAL characters.

According to this author's knowledge, no existing database has aimed to obtain listener vocalizations for synthesis, as the objectives of available databases are entirely different.

6.3 Proposed method for data collection

Traditionally, speech synthesis databases, including expressive speech material, are recorded in a studio environment with a single speaker using predefined recording scripts. However, listener vocalizations appear unnatural with predefined recording scripts. Therefore, the traditional approach of recording setup is not useful to capture listener vocalizations.

Listener vocalizations seem to appear natural only in conversation. We propose a method to record a natural dialog speech between an actor and his dialog partner in an anechoic studio. The actor is the same person with whom we had recorded expressive speech synthesis databases in the past. In the new proposed recording setup, shown in Figure 6.1, the actor and his dialog partner sit in different rooms and hear each other using headphones, so that we can record each speaker's voice on a different channel without interference of the other speaker's speech. The conversation between the speakers follows a general *chatting* scenario, where the dialogue has no predefined scripts. In order to maintain natural conversation, we record dialogue only in sessions of 20 minutes each. As we are aiming to capture listener vocalizations, the actor will be instructed to participate in a free dialog, but to take predominantly a listener role.

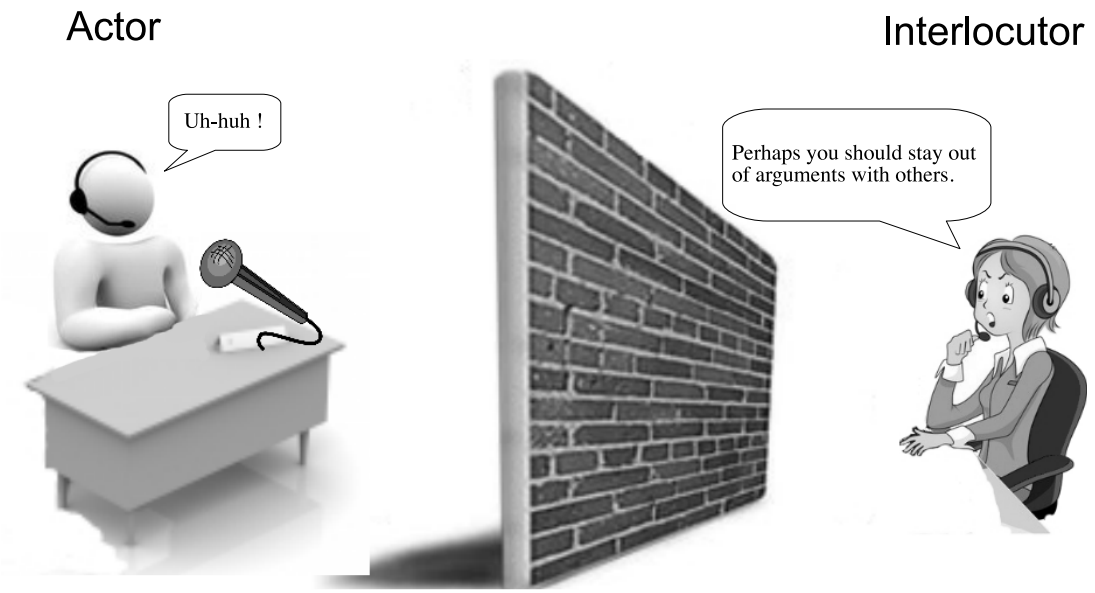


Figure 6.1: A schema of the recording set-up

6.4 Experimental collection of German data

In preparation of the recordings of the SAL voices, we experimented with a German actor who pretended to be each of the SAL characters in turn while chatting with a human partner. In the first instance, we applied the proposed methodology described in the previous section to record German listener vocalizations. We recorded dialog speech in a studio environment to get a good quality and anechoic speech corpus.

Our speaker was a professional male German actor with whom we had recorded expressive speech synthesis databases in the PAVOQUE[1] project (Steiner et al. 2010). He is also the same actor with whom we had recorded a poker style database for the project IDEAS4Games (Schröder et al. 2008a; Gebhard et al. 2008), an emotion-aware poker game, in which two agents and a user played against each other with physical cards carrying RFID tags. Using the same speaker (actor) was essential for being able to use the recorded vocalizations with our synthesis voices in the future.

The actor was instructed to participate in a free dialogue, but to take predominantly a listener role. We encouraged him to use "small sounds that are not words", such as *mm-hm*, where it felt natural, in order to keep his interlocutor talking for as long as

[1] http://mary.dfki.de/pavoque

possible. However, he was also allowed to "say something" and therefore to become the speaker in the conversation where this "felt natural" to keep the dialogue going.

Recordings were made in several different stages. In the initial stage, we instructed the actor to "be himself" (not to act) and in the later stages, he was instructed to act like one of three characters representing different emotionally colored personalities (Douglas-Cowie et al. 2008): Spike is always aggressive, Obadiah is always gloomy, Poppy is always happy. These characters have been designed to represent different quadrants of the arousal-valence plane, and the actor was acquainted with their definitions from previous recordings.

Sessions lasted about 20 minutes each. Their durations vary slightly according to the actor's ability to maintain a consistent personality during the conversation.

Two female student assistants, one of whom had worked with the same actor in the past, took turns as the dialog partner, talking to him about various emotionally loaded topics of their choice. The dialogue partners were sitting in separate rooms, but they could see each other through a glass wall and hear each other using headphones, which enabled an audio-visual interaction. Each speaker's voice was recorded on a separate channel. We also recorded the actor's face using a standard MiniDV camera, enabling future study of audio-visual synchrony in listener behavior. In this thesis, only the analysis of the audio data is reported. We used a simple audio-visual recording set-up (a Mini-DV camera with PAL resolution and 25 fps, stereo sound at 32 kHz; the actor was recorded on the left channel, the partner on the right channel), in order to have audiovisual models of integrated non-verbal backchannels in view of joint synthesis in face and voice.

Overview of German listener vocalizations

As a result of the database collection exercise, we obtained around six hours of German dialog speech. Listener vocalizations were identified and marked on the time axis by our student assistants. Only the actor's listener vocalizations are being used. Table 6.1 shows the German material used in this thesis.

The actor status	Corpus duration (in minutes)	Number of listener vocalizations
Natural	190	568
Obadiah	45	181
Poppy	45	93
Spike	70	238
Total	350	1080

Table 6.1: The number of listener vocalizations obtained when the actor is being himself (natural) or acted like an emotional character.

Single-word description

At first sight, when we look at the interactive speech corpus, we observed that the dialogue speech contains not only verbal vocalizations but also many non-linguistic vocalizations. Different types of non-linguistic vocalizations like *laughter* and *sigh* were observed. A single-word description annotation schema was proposed to annotate this kind of data, where we aim to get a simple description (i.e. mostly in a single word) for each listener vocalization. The descriptions would be usually one word descriptions and exceptionally multi-word descriptions. In order to annotate non-linguistic vocalizations as well, we also instructed annotators that they could use para-language descriptors such as *(laughter)* and *(sigh)*. Although it is an open-ended set of descriptors, the starting point for the set was six para-language descriptors provided by Douglas-Cowie, Cowie, and Schröder (2003): *(laughter), (sobbing), (gasp), (sigh), (snort)* and *(scream)*.

The corpus was annotated by the two student assistants according to the proposed schema using Praat software (Boersma and Weenink 2010). We annotated only vocalizations produced by our target speaker, not of the interaction partner. Based on descriptions provided by student assistants, Table 6.2 shows the frequency of response tokens used by the actor. The table shows only descriptions found at least thrice in the corpus. Among 1080 listener vocalizations uttered by the actor, the first four frequent response tokens, *mhm, laughter, ja* and *sigh*, show more than 50% of coverage.

When the actor impersonated different emotional characters, he also varied enormously with respect to the type of listener vocalizations uttered (see Figure 6.2). While

Description	Number of occurrences	Description	Number of occurrences
mhm	204	richtig	7
(laughter)	164	achso	6
ja	130	gut	5
(sigh)	57	O.K.	5
mh	56	okay	5
ah	39	(gasp)	4
mhmh	32	jo	4
aha	28	mm	4
mja	21	und	4
oh	14	(laughter)_ja	3
(snort)	13	ach	3
hm	13	ahja	3
mmm	12	echt	3
hmm	11	ey	3
mhm_mhm	9	hhmhm	3
(sigh)_mhm	8	ja_(laughter)	3
genau	7	mh_mh	3
nein	7		

Table 6.2: The description and its number of occurrences of listener vocalizations obtained from the German corpus

ja, mhmh and *(laughter)* are predominantly used response tokens in Poppy's vocalizations, the set of natural (non-acted) vocalizations contains *mhm, (laughter)* and *yeah* vocalizations. Whereas the actor being Spike uttered *(laughter), ja*, and *aha* frequently, Obadiah used *mhm, (sigh)*, and *ja* response tokens. However, *mhm*, and *ja* are frequent tokens used in all roles.

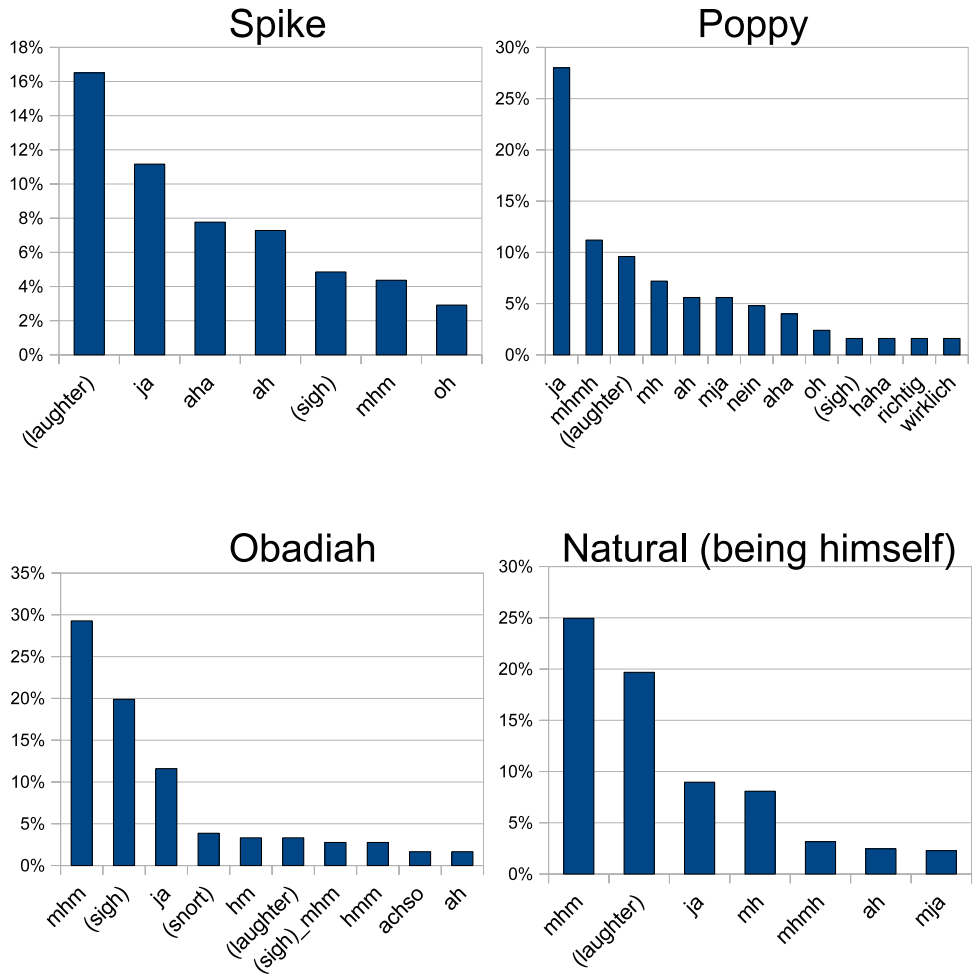


Figure 6.2: Most frequent tokens used when the actor was "being himself" (natural) or acted like an emotional character

6.5 Collection of British English vocalizations

In a second data collection task, new British English speech synthesis voices were created for the four SAL characters described in Chapter 1. For each character, a professional actor was selected based on the voices that seemed to fit best with the facial model and the intended personality of the character. Recordings were carried out in an anechoic chamber throughout the course of one week. Even though the British actors were originally chosen just for the recordings required for building new TTS voices, they were also asked to record free dialogue of around 30 minutes. The experience gained with the German data collection were used to work with the British English voices recorded for the SAL characters in a more structured way. Here, a different speaker produced the speech material for each of the four SAL characters.

The approach and instructions were almost the same as the data collection procedure used for the German corpus. Actor and interlocutor were located in different rooms so that their voices could be recorded onto separate audio channels; the actor heard the interlocutor through closed system headphones, to avoid leakage of the headphone output to the actor's microphone. Actors were instructed to participate in a free dialogue, but to take predominantly a listener role. We encouraged them to use, "small sounds that are not words", such as *mm-hm*, where it felt natural, in order to keep their interlocutor talking for as long as possible. However, they were also allowed to, "say something" and therefore to become the speaker in the conversation where this, "felt natural" to keep the dialogue going. One of three experimenters acted as the interlocutor.

	Prudence	Poppy	Spike	Obadiah	Total
Corpus duration (in minutes)	25	30	32	26	113
Number of vocalizations	128	174	94	45	441

Table 6.3: British English listener vocalizations recorded for the four SAL characters

Spike		**Poppy**	
Description	Number of occurrences	Description	Number of occurrences
yeah	22	yeah	61
(laughter)	20	(sigh)	26
mhm	9	(laughter)	14
aha	4	mhmh	7
hm	3	(gasp)	6
(snort)	2	oh	6
absolutely	2	gosh	4
ya	2	mh	4
well	2	definitely	4
		yes	4
		really	3
		oh (laughter)	3
		yeah (laughter)	3
		wow	2
		really (laughter)	2
		no	2
		alright	2

Obadiah		**Prudence**	
Description	Number of occurrences	Description	Number of occurrences
mhmh	8	yes	31
(sigh)	8	yeah	23
yeah	7	right	18
right	6	tsright	7
(laughter)	3	aha	6
alright	2	tsyeah	6
		(laughter)	6
		mhm	3
		tsalright	3
		gosh	3
		tsyes	3
		alright	2
		really	2
		yes (laughter)	2

Table 6.4: Descriptions of British English listener vocalizations recorded for all the four SAL characters: Poppy (cheerful), Prudence (pragmatic), Spike (aggressive) and Obadiah (gloomy)

Once the dialogue speech for all four characters was recorded, listener vocalizations were marked on the time axis and transcribed as a single (pseudo-)word, such as *myeah* or *(laughter)*. The speakers varied enormously with respect to the number of listener vocalizations produced. Whereas Obadiah produced only 45 vocalizations, Poppy produced 174 (see Table 6.3). The table also shows the most frequent tokens used by each character.

The single-word description was also applied to British English corpus. The descriptions of British English listener vocalizations recorded for all the four SAL characters are shown in Table 6.4. The table shows the descriptions of listener vocalizations that are uttered by each speaker at least twice. The coverage of descriptions is high for Poppy and Prudence, whereas it is low for Spike and Obadiah.

The speakers also varied enormously with respect to the type of listener vocalizations uttered. While *yeah, (sigh)* and *(laughter)* are predominantly used response tokens in Poppy's vocalizations, Prudence produced *yes, yeah*, and *right* vocalizations. Whereas Spike uttered *yeah, (laughter), mhm* frequently, Obadiah used *mhmh, sigh, yeah* response tokens. However, *yeah* is the most frequent token used among all speakers.

6.6 Summary

This chapter has presented a data collection of listener vocalizations in interactive speech synthesis. We have described a method for collecting listener vocalizations in view of emotionally colored conversational speech synthesis. This method has considered generic and project specific requirements for better quality of speech synthesis and the possibility of implementing feedback mechanisms. It has been applied for two different languages: German and British English. Whereas German vocalizations were gathered from a single speaker, British English vocalizations were collected from four different professional British actors. A single-word description annotation schema has been applied to the extracted vocalizations from dialogue recordings, which enables better insight to the available listener vocalizations in this corpus. Although the variety of vocalizations available in each of the corpora is limited, the proposed method appears to be successful for the collection of natural listener vocalizations.

Chapter 7

Exploratory annotation

The quality of corpus driven speech synthesis technologies depends on the quality of annotation of the speech corpus. As described in previous chapters, traditional speech synthesis systems annotate speech material as a sequence of phonetic symbols. Most of the process happens automatically using force-alignment techniques. When we come to listener vocalizations, however, the meaning of vocalizations depends on several behavior (form) properties. Such behavior properties include not only the phonetic sequence (i.e. segmental form) of the speech but also prosodic parameters like intonation and voice quality.

The annotation of vocalizations, as described in Chapter 5, is required on two different levels: meaning annotation and behavior annotation. This chapter presents our exploratory approach to investigate the annotation of meaning and behavior of listener vocalizations. We also describe our investigation on the following research questions:

- What are suitable meaning and behavior descriptors for listener vocalizations?
- How to annotate meaning and behavior of a listener vocalization?

The annotation of listener vocalizations in our data progressed in several stages. In Section 7.1, we present the data used in this exploratory annotation study. During an initial screening process, listener vocalizations were identified, their occurrences were marked on the time axis, and a simple initial coarse description of meaning and behavior was carried out using an "ABL" annotation scheme (see Section 7.2). In a second stage, as described in Section 7.3, a fuller analysis was carried out by means

of detailed, informal descriptions of each listener vocalization. Based on such descriptions, we find suitable meaning descriptors to represent vocalizations (see Section 7.4). In the later stages, the full descriptions of meaning and behavior were summarized in terms of behavior categories and meaning categories associated with types of reference, as described Section 7.5. In Section 7.6, we describe our approaches to annotate behavioral properties such as intonation and voice quality. The corpus was annotated by the same two student assistants using Praat (Boersma and Weenink 2010). Finally, Section 7.8 summarizes our experience with this whole exercise.

7.1 Corpus used for investigation

From the German recordings, as described in Chapter 6, we obtained six hours of German dialog speech. A sub-corpus of around five hours was used for the three-stage exploratory annotation described in this Chapter. Only the actor's listener vocalizations are being used. Table 7.1 shows the material used in this investigation.

The actor status	Corpus duration (in minutes)	Number of listener vocalizations
Natural	190	568
Obadiah	45	181
Poppy	45	93
Spike	30	125
Total	310	967

Table 7.1: Corpus duration in minutes when the actor is being himself (natural) or acted like an emotional character.

7.2 ABL scheme

From the first sight of the corpus, we observed that many of the listener vocalizations could be characterized in terms of three overlapping categories: +/- affect, +/- backchannel and +/- laughter. Different combinations were observed, such as affective

backchannels, laughter as backchannels and affective laughters. Therefore, an ABL annotation scheme was used, where A stands for Affect, B stands for Backchannel and L for Laughter, and each can be present or absent. For example, the annotated tag 'AL' tells that the corresponding vocalization is laughter and it carries affective meaning, but it is not a backchannel. According to this scheme, the annotators had to identify listener vocalizations, mark the occurrence on the time axis, and then provide an 'ABL' tag. For the annotation, backchannels were operationalized as short utterances like *mm-hm* and *uh-huh* which appeared to encourage the speaker to continue talking.

Analysis

The annotation of 967 listener vocalizations according to the ABL annotation scheme was provided in the first phase. Among all listener vocalizations, 51.5% were labelled as affective, 75.5% as backchannel and 20% as laughter. The distribution of A, B and L is shown in Figure 7.1. Among the backchannels, 29.6% were labeled as affective (i.e., A+B or A+B+L), which means that more than one third of vocalizations with backchannel function were also transmitting affective meaning. Most of laughter was labelled either as a backchannel or as affective or both.

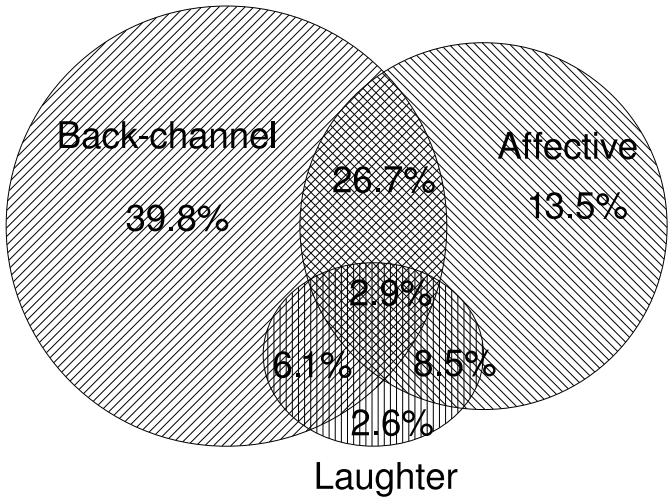


Figure 7.1: The distribution of listener vocalizations according to ABL annotation scheme

7.3 Informal descriptions

In order to get a clearer picture of the data, we used a detailed informal description of each vocalization before trying to find suitable categories to represent the meaning and behavior observed. Subsequent grouping of these descriptions will help to understand the types of form and meaning of listener vocalizations, at least for the speaker we studied. Although the annotation of a detailed informal description for each listener vocalization is a time consuming process, we wanted to make sure that we are not blinded by looking through the pattern of a pre-existing set of categories. Therefore, we had the content, form and subtexts of each listener vocalization annotated with informal descriptions in the annotator's own words, as shown in Figure 7.2. The form provides information about phonetic segments, voice quality, duration and/or intonation. Similarly, the content and "subtext" tiers describe the meaning and, optionally, a suitable text substitution.

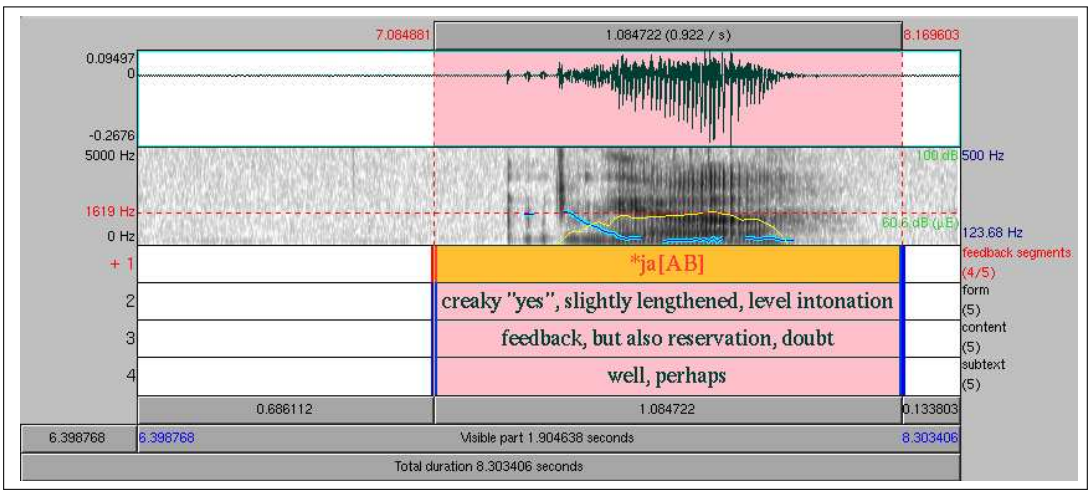


Figure 7.2: Example of an informal description for a listener vocalization, where the first tier represents annotation according to the ABL scheme, the second tier represents form, the third tier content and the fourth tier subtext.

7.4 Sources of meaning descriptors

In order to abstract away from the detailed, individual descriptions towards a generalized summary view of the meaning conveyed in our data, we used a categorical annotation. Based on the informal descriptions, we aimed for a limited set of categories that capture the essence of the meaning as recorded in the descriptions. We considered it important for the initial informal descriptions not to be guided by any pre-existing framework. It seems appropriate for the consideration into categories to attempt using an existing set of meaning categories from the literature, and to verify to what extent it covers the meaning contained in our data. In this section, we describe such existing set of descriptors for meaning and reference annotation.

7.4.1 Baron-Cohen's epistemic states

Baron-Cohen et al. (2001) developed a set of epistemic mental states (see Table 7.2) in his research on children with Autism Spectrum Disorder (ASD). ASD is a developmental disorder that is characterized by impairments in reciprocal social interactions and relationships, verbal and nonverbal communication, unusual and repetitive behaviors. They argued that children with autism need help in understanding that others have mental states similar to or different from their own mental ideas. They also have developed software for such education (mind-reading) and an animation series to teach children with autism to recognize and understand emotions.

The epistemic states developed by Baron-Cohen et al. (2001) have become popular in other fields such as social interaction. In human-human conversation, the listener assumes that the meaning of an utterance will be relevant to the speaker's current intentions and vice versa.

In order to capture the essence of meaning or intention of informal descriptions, we can expect that the set of epistemic states to be a good starting point.

Category Number	Baron-Cohen's Epistemic states	Meaning
1	ACCUSING	blaming
2	ANTICIPATING	expecting
3	CAUTIOUS	careful, wary
4	CONCERNED	worried, troubled
5	CONFIDENT	self-assured, believing in oneself
6	CONTEMPLATIVE	reflective, thoughtful, considering
7	DECISIVE	already made your mind up
8	DEFIANT	insolent, bold, don't care what anyone else thinks
9	DESIRE	passion, lust, longing for
10	DESPONDENT	gloomy, despairing, without hope
11	DISTRUSTFUL	suspicious, doubtful, wary
12	DOUBTFUL	dubious, suspicious, not really believing
13	FANTASIZING	daydreaming
14	FLIRTATIOUS	brazen, saucy, teasing, playful
15	FRIENDLY	sociable, amiable
16	HOSTILE	unfriendly
17	INSISTING	demanding, persisting, maintaining
18	INTERESTED	inquiring, curious
19	NERVOUS	apprehensive, tense
20	PANICKED	distraught, feeling of terror or anxiety
21	PENSIVE	thinking about something slightly worrying
22	PLAYFUL	full of high spirits and fun
23	PREOCCUPIED	absorbed, engrossed in one's own thoughts
24	REFLECTIVE	contemplative, thoughtful
25	REGRETFUL	sorry
26	SCEPTICAL	doubtful, suspicious, mistrusting
27	SERIOUS	solemn, grave
28	SUSPICIOUS	disbelieving, suspecting, doubting
29	TENTATIVE	hesitant, uncertain, cautious
30	THOUGHTFUL	thinking about something
31	UNEASY	unsettled, apprehensive, troubled
32	UPSET	agitated, worried, uneasy
33	WORRIED	anxious, fretful, troubled

Table 7.2: Baron-Cohen's Epistemic states described in (Baron-Cohen et al. 2001)

7.4.2 Geneva emotion wheel categories

While Baron-Cohen categories include epistemic mental states, they do not seem to represent strong emotions. From the informal descriptions, we find that some vocalizations have strong emotional intention.

To measure subjective feelings, Scherer (2005) offers us an efficient, simple method with the Geneva Emotion Wheel (See Figure 7.3). The Geneva Emotion Wheel (Scherer 2005) was devised as a tool for the verbal report of emotions. It includes 16 emotion categories positioned in a circle. The 16 categories are ordered according to their postulated position in a 2 dimensional space. The two underlying dimensions are the level of perceived control in the situation that generates the emotion (vertical dimension) and the positive/negative (pleasant/unpleasant) quality of the situation and of the resulting feeling (horizontal dimension).

7.4.3 Bühler's Organon model

In addition to the annotation of meaning as such, it became apparent from our informal descriptions that several kinds of reference should be distinguished. Indeed, listener vocalizations seemed to differ with respect to their reference: is the listener providing information about his own internal state (self expression), is he reaffirming the relationship with the speaker (stance towards the other), or is he commenting about the current topic of discussion (attitude towards the topic)?

Bühler's (Bühler 1934) Organon model (Figure 7.4) provides a structure distinguishing these three types of reference of an expression. In his terms, a "symptom" has the function of *expression* of the sender's state; a "signal" serves as *appeal* to a receiver; and a "symbol" is used as a *representation* of objects and facts. According to Bühler, all three functions are co-present in spoken communication, though their relative salience can vary. In our terms, this suggests we should distinguish a *self reference* (in which our listener expresses his own state), a reference towards the *other* (where the vocalization is used to signal the listener's stance towards the speaker), and a reference towards the *topic*. Following Bühler, all three functions can be expected to be co-present.

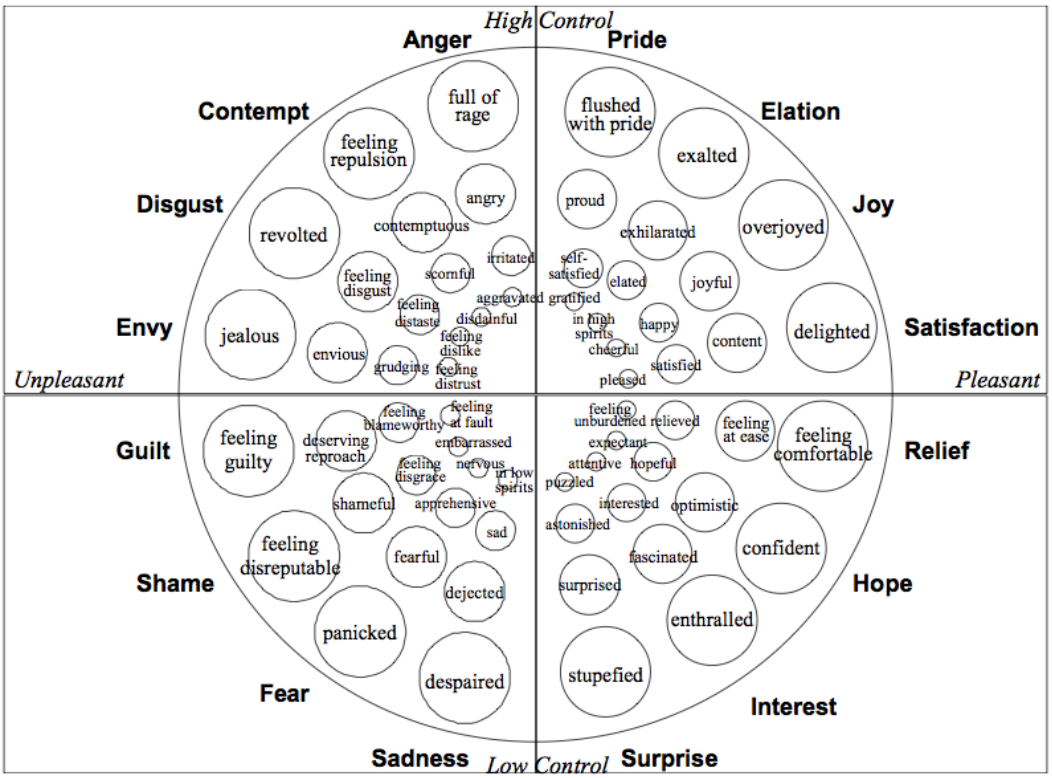


Figure 7.3: Geneva Emotion Wheel

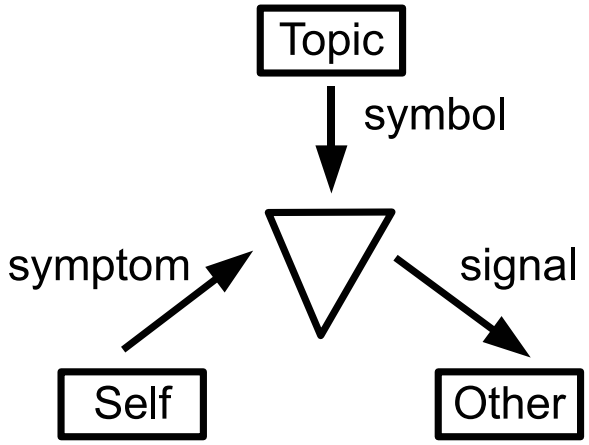


Figure 7.4: Bühler's Organon model of speech, adapted from (Scherer 1988).

7.5 Categorical meaning annotation

7.5.1 Procedure

We used the Baron-Cohen (Baron-Cohen et al. 2004) set of 33 categories describing epistemic-affective states as a starting point for our tag set. Annotators were instructed to use only those categories from the set that seemed appropriate, and to add categories that seemed necessary to describe the data but were not contained in the Baron-Cohen set. They could use categories from the Geneva Emotion Wheel (Scherer 2005) or propose their own category labels as they felt appropriate. No restrictions were made concerning the minimum or maximum number of categories to use. The same annotators who wrote the informal descriptions also assigned the categories, based on the informal descriptions and the recordings.

Annotators were instructed to provide a categorical annotation as follows. For any given listener vocalization, they had to provide at least one category; where the expressed meaning seemed too complex to be covered by a single category, they could use up to three categories. For each category used, they could optionally indicate the reference according to the Organon model: (S)elf reference, (O)ther reference, or (T)opic reference.

7.5.2 Results

Meaning categories

Annotators used 24 out of the 33 Baron-Cohen categories. They added nine out of the 40 categories of the emotion wheel (Scherer 2005), as well as four custom categories. The 37 categories used are shown in Table 7.3. The number of frequently used categories is much smaller, though. Only five categories were used on at least 10% of the vocalizations, and eleven categories were used on at least 5% of the data.

Annotators made frequent use of the possibility to give more than one category. 17.7% of the vocalizations were labelled with a single category; 52.9% were labelled with two categories, and 29.4% with three categories.

The characters clearly differed with respect to the categories of meaning conveyed by their listener vocalizations. In his “natural” interaction mode, the actor is friendly,

Baron-Cohen categories	**anticipating**, cautious, concerned, confident, contemplative, decisive, defiant, **despondent**, **doubtful**, **friendly**, hostile, insisting, **interested**, nervous, playful, preoccupied, regretful, serious, suspicious, **tentative**, **thoughtful**, uneasy, upset, worried
Emotion wheel categories	**amused**, angry, compassionate, disgusted, happy, **irritated**, relieved, **scornful**, **surprised**
Custom categories	depressed, excited, ironic, outraged

Table 7.3: The list of categories used for annotation. Frequently used categories ($>$ 5%) are highlighted in bold, and most frequent categories ($>$ 10%) are underlined.

interested and amused; as Spike, he is scornful, irritated, amused and ironic; as Obadiah, he is despondent and friendly; and as Poppy, he is interested and friendly (see Figure 7.5). This seems partly but not fully consistent with the intended personalities. A more fine-grained analysis taking into account reference annotation in addition to these meaning categories seems to show a clearer picture (see below).

Reference types

Annotators made very frequent use of the reference types in annotation. In 31% of the cases, they actually used all three references, which means that they considered self-related, other-related and topic-related meaning to be present in a single vocalization. In 48% of the cases, two reference types were indicated (i.e., S+O, O+T or S+T). In 14.3% of the cases, only one reference was given, and in 6.7%, no reference was specified.

The Self, Other and Topic reference based distinction seems to provide insights in the characters' expressive behavior, as shown in Figure 7.5. For example, the optimistic character (Poppy) shows mostly happy self expression, he is interested in the Topic, while being friendly and compassionate towards the Other.

Indeed, self-expression seems to describe very well the intended personality: despondent, irritated, uneasy and thoughtful for Obadiah, the gloomy character; happy, interested, surprised, thoughtful, excited and amused for the cheerful character Poppy; and for the aggressive character, Spike, self-expression is amused, irritated, ironic,

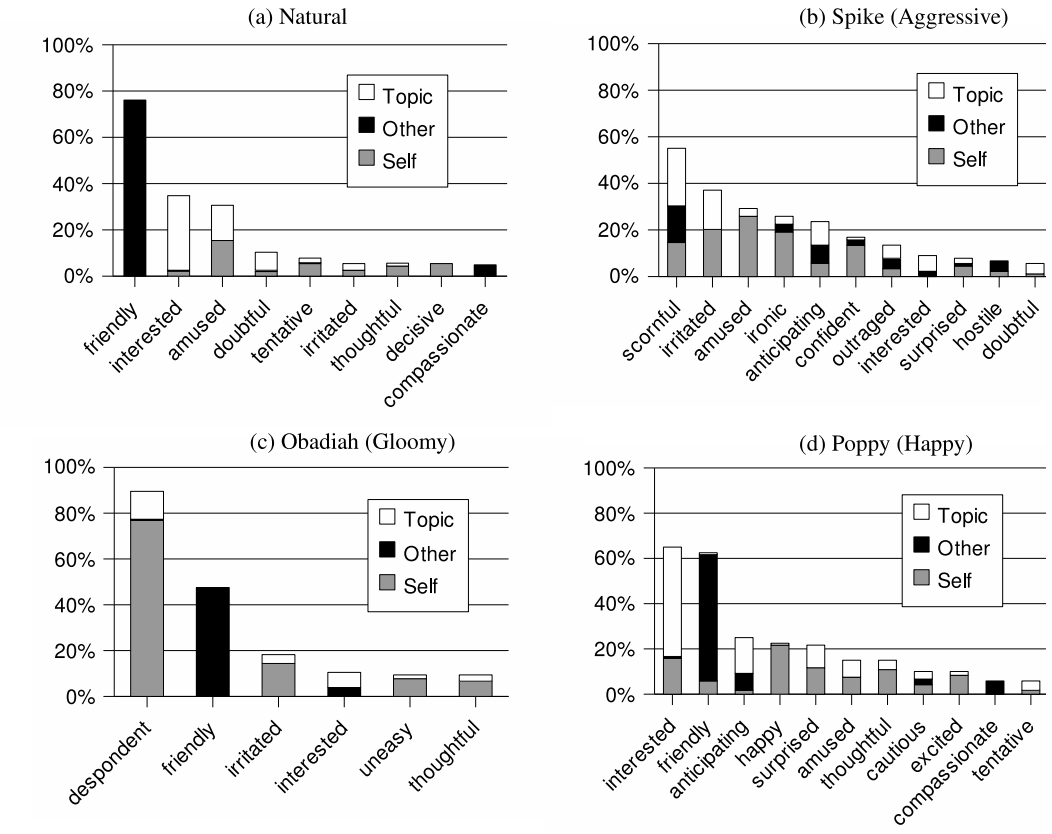


Figure 7.5: Most frequent affective-epistemic categories and associated reference types, per character.

scornful, and confident. In the same way, we can now characterize the "natural" speaking mode of our actor as amused, sometimes decisive and sometimes tentative, and thoughtful.

The only category that does not quite seem to fit the picture is the observation that Spike is predominantly amused. To understand better the instances in which Spike is amused, we show the most frequent categories co-occurring with "amused" for Spike and for the other character showing substantial self-amusement, the natural speaking mode of the actor (Figure 7.6). It is very obvious that Spike's amusement co-occurs nearly exclusively with negative emotions such as scornful, outraged and ironic, whereas the natural actor shows amusement mostly with the positive categories friendly and interested. This suggests that the two kinds of amusement are actually very different – a point that would have been difficult to make if only a single meaning

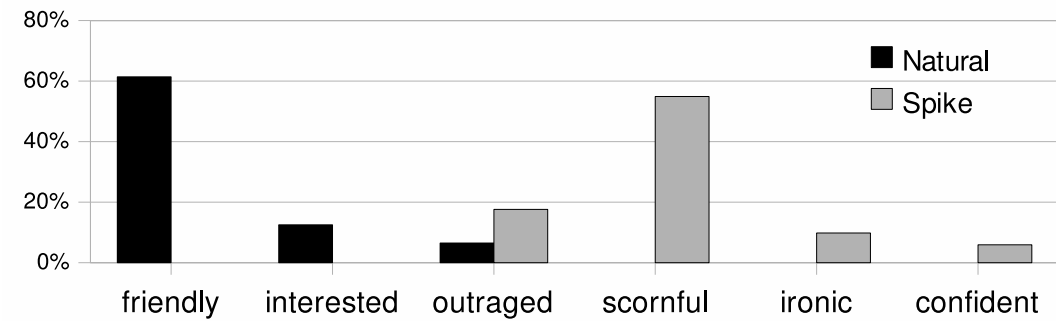


Figure 7.6: Most frequent ($> 5\%$) meaning categories co-occurring with the category 'amused', for natural and Spike listener modes.

category had been annotated.

The Other reference seems to show clear differences in interpersonal stance among the characters. For Spike, the aggressive character, Other-related expressions are scornful, outraged, ironic or hostile, whereas other characters are friendly or compassionate. The attitude towards the topic of discussion seems to be sensibly indicated by the Topic reference: the actor himself and Poppy show a lot of interest, whereas Spike shows a predominantly scornful and irritated attitude, and Obadiah shows little topic-related signs at all.

These results suggest that distinguishing the reference in addition to affective-epistemic meaning categories may be a useful means to gain insights regarding a character's mood or personality (Self reference), interpersonal stance (Other reference) and attitude towards a topic (Topic reference).

7.5.3 Inter-rater agreement

A subset of 102 listener vocalizations from the non-acted part of the dialog corpus was annotated by both annotators with meaning and reference categories as described above. As we allowed for more than one category per instance, we computed Cohen's Kappa separately for each category, treating annotations as a binary "present/absent" feature. On this basis, we computed Kappa for each meaning category and each reference type.

As shown in Figure 7.7, the Kappa values for the most frequently used meaning categories friendly, interested and amused were 0.02, 0.41 and 0.82 respectively. Among

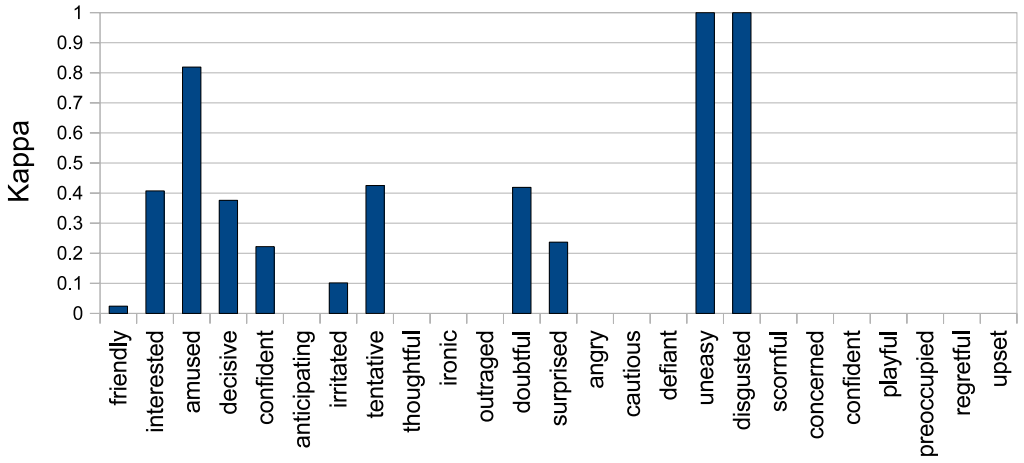


Figure 7.7: Inter-rater agreement of meaning categories

the less frequent categories, Kappa values for decisive, confident, tentative, doubtful and surprised scores range between 0.22 and 0.43, whereas anticipating, thoughtful, ironic, irritated, outraged, angry show nearly no agreement between two annotators.

For reference categories S, O, and T, Kappa was close to 0, indicating no consistent agreement between the two annotators. It remains to be seen whether this is due to an intrinsic ambiguity or due to insufficient instructions.

The agreement in the meaning annotation provided by the raters is very low. Possible reasons, we could think of, are: (i) the *open-endedness* of the task that might have created confusion among the annotators; and (ii) the size of the meaning descriptors list is large.

7.6 Behavior annotation

As described in Chapter 2, the term 'behavior' is refers to acoustic properties like segmental form and prosody in the context of this thesis. Although the relation between meaning and behavior is not deeply explored yet, the literature (in Chapter 2) argues that segmental form and prosody of vocalizations have significant impact on meaning. During initial coarse description, as described in Chapter 6, vocalizations were annotated with the segmental form. In addition, this section describes our efforts for

phonetic segmentation of vocalizations. This section also describes our methodology to find suitable descriptors for intonation and voice quality.

7.6.1 Phonetic alignment

Phonetic alignment of speech is always required for ECA's lip synchronization (Chapter 4). Hand-labelled phonetic segment labels for all vocalizations were provided by a phonetically trained student assistant. The manual labels of a vocalization contain time-stamps of each phonetic segment as well as corresponding suitable phone description. This is suitable for vocalizations with a phonemic structure such as *myeah*, but is problematic for other vocalizations such as *laughter, sighs*, or *a rapid intake of breath*. In these cases, the viseme-based mouth shapes can only serve as coarse approximations of natural behavior.

7.6.2 Intonation

The pitch tracking algorithms, nowadays, are computing pitch contours reasonably well. In this section, we propose and experiment with a semi-automatic procedure to annotate intonation contours.

Firstly, an intonation contour can be automatically computed by fitting a polynomial to f0 values extracted using a pitch tracker; because polynomials can approximate intonation contours of speech signal, and they can handle problems with unvoiced regions and large pitch excursions. Secondly, separately for each speaker, we use unsupervised clustering of intonation contours to identify the vocalizations with a similar prosody. We make sure that the distance measure used to cluster contours should consider the shape of the contour instead of contour mean height. Finally, we find similar clusters and name them with suitable labels.

Polynomial fitting

For each listener vocalization, as described above, polynomial coefficients can be computed on f0 values extracted using the pitch tracker Snack (Sjölander 2006). Figure 7.8 shows different polynomial orders to approximate an intonation contour. In this example, the second and third order polynomials preserve the shape of intonation contour.

After analyzing many intonation contours, we concluded that the second order polynomial would be a reasonable trade off between simple approximation and preserving contour shapes.

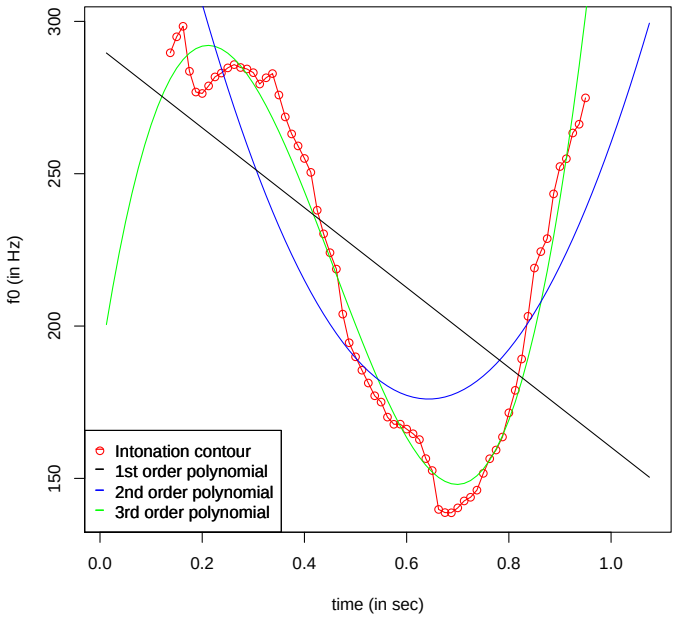


Figure 7.8: First, second and third order polynomial fitting to an intonation contour

Distance computation

In order to cluster similar-looking pitch contours, we look for a meaningful approach to compute distances between pitch contour pairs. We have to cluster pitch contours that have a similar shape, so differences in pitch height or pitch range should not be included in the distance measure. Equation 7.1 states that one minus the Pearson product moment correlation is used to calculate the distance between each pair of pitch contours. This distance metric is similar to calculating distances between z-normalized pitch contours, subtracting their mean value and dividing them by their standard deviation. By the given distance metric, the clustering procedure considers intonation contour shape but not the differences in pitch height or range.

$$\begin{aligned} D &= 1 - corr(F_{0i}, F_{0j}) \\ &= 1 - \left(\frac{1}{n-1} \sum \left(\frac{F_{0i} - \tilde{F}_{0i}}{sd_{F_{0i}}} \right) \left(\frac{F_{0j} - \tilde{F}_{0j}}{sd_{F_{0j}}} \right) \right) \end{aligned} \tag{7.1}$$

Unsupervised clustering

We use Hierarchical agglomerative clustering (HAC) algorithm to cluster intonation contours with similar shape. Agglomerative techniques are commonly used for unsupervised clustering. Hierarchical agglomerative cluster analysis is a statistical approach for finding relatively homogeneous clusters of data objects based on measured characteristics. The algorithm starts with every single object in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster. The algorithm of the agglomerative hierarchical clustering method, according to Härdle and Simar (2007), is described in Listing 7.1.

```
1. Construct the finest partition
2. Compute the distance matrix
3. DO
        – Find the clusters with the closest distance
        – Put those two clusters into one cluster
        – Compute the distances between the new groups
          and the remaining groups
4. UNTIL all clusters are agglomerated into one group
```

Listing 7.1: Agglomerative hierarchical clustering algorithm (Härdle and Simar 2007)

HAC implementation can be done in several ways such as single-linkage, complete-linkage and average-linkage clustering. For each possible way, the methodology used to compute the distance between clusters is different. In single-linkage clustering, HAC algorithm considers the shortest distance between two clusters while computing distance matrix. On the other hand, the algorithm considers the longest distance between clusters in complete-linkage clustering. In case of average-linkage clustering, the average distance between each cluster pair will be considered.

Analysis of intonation annotation

As described in the previous section, we used a HAC algorithm to cluster similar shaped intonation contours. The complete linkage HAC is used in order to make inter-cluster distance is as high as possible. To make sure that the clustering algorithm

weight on contour shapes instead of mean height of the contour, the pearson product moment correlation is used as all pair-wise distances between intonation contours.

F0 Descriptors
low, mid, high, rising, falling, rise-fall, and fall-rise

Table 7.4: The final set of descriptors used for intonation annotation

Once we cluster all intonation contours into K clusters, we manually assign a suitable label to each cluster using a visualization tool which shows the polynomial contours of the cluster. The resulting labels for describing intonation contours are: *level, rising, falling, rise-fall*, and *fall-rise*. We then subdivided *level* intonation contours into three different level contours – *low, mid* and *high* level contours – based on different threshold settings for height of the contour mean value. This semi-automatic procedure seems to provide the better labeling accuracy with lesser manual efforts. Table 7.4 shows the final set of descriptors used for intonation annotation. The intonation contours are clustered as shown in Figure 7.9.

7.6.3 Voice quality

In contrast to the intonation annotation, voice quality can not be annotated with an automatic procedure. In order to find suitable descriptors for manual annotation of voice quality, we have gone through informal descriptions of vocalizations again. The behavior (form) level informal descriptions, see Figure 7.2, contain different voice quality parameters like creaky, breathy and modal. Inspired by voice quality descriptions used in informal descriptions, Laver (1991)'s voice quality categories (see Table 7.5) seem suitable to describe voice quality of listener vocalizations.

Considering the possibility of multiple voice quality parameters in a single vocalization, we allowed annotators to use multiple categories (maximum up to three) for each listener vocalization where that seemed appropriate, but they have to prioritize the categories in the voice quality annotation. According to instructions given to annotators, the primary label given to each vocalization is the strongest voice quality component in the vocalizations. For example, if the voice quality description provided for a vocalization is “creaky, breathy", the vocalization is more *creaky* than *breathy*.

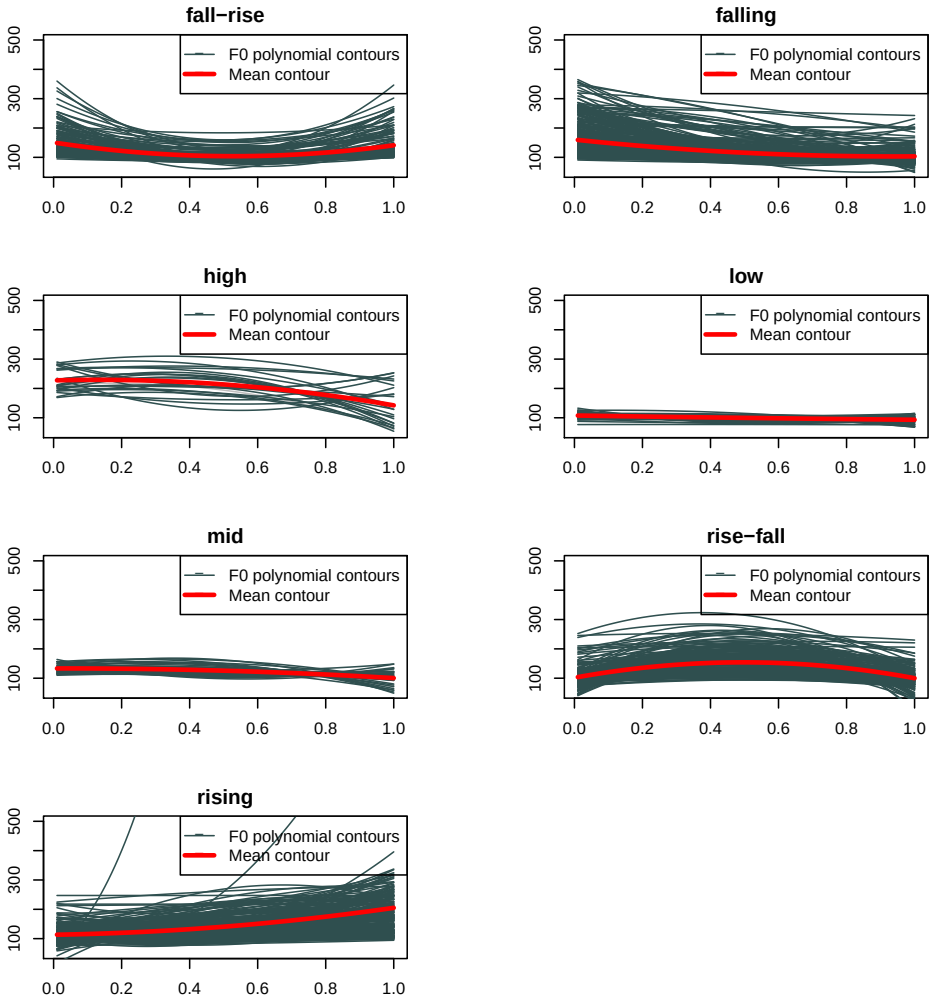


Figure 7.9: Clusters of intonation contours

Therefore, the first label is the primary category and the second and third labels are the secondary categorical descriptions.

Analysis of voice quality annotation

The voice quality annotation was provided by a student assistant according to instructions given in the previous section. This section describes the analysis of voice quality annotation.

Among all listener vocalizations, 78% of listener vocalizations are annotated with a single voice quality category. The remaining vocalizations are annotated with two

modal	neutral mode of phonation where no specific feature is explicitly changed or added
creaky	a train of discrete laryngeal pulses; phonation with strong adductive tension and low frequency
whisper/whispery	low volume voice with low tension and audible friction of air in and above larynx (used to show secrecy and intimacy)
breathy	auditory impression very similar to whispery voice but produced with less laryngeal effort and less audible friction (used to express intimacy, relaxation and satisfaction)
tense	ligamental, harsh or ventricular phonation; louder and higher-pitched
lax	breathy or whispery phonation; softer and lower-pitched

Table 7.5: Voice quality labels (Definitions are taken from Laver 1991)

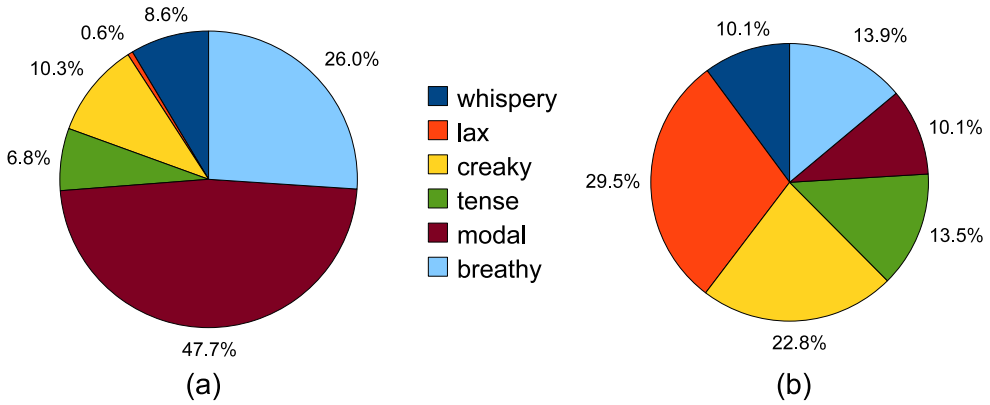


Figure 7.10: The distribution of listener vocalizations according to voice quality annotation: (a) primary categorical descriptions (b) secondary categorical descriptions

categories, but no vocalization was annotated with a third label. Figure 7.10 shows the distribution of both primary and secondary voice quality descriptions. The most frequent primary categories seem to be *modal* (47.7%), *breathy* (26%) and *creaky* (10.3%), whereas *lax* (29.5%) and *creaky* (22.8%) are the most frequent secondary voice quality categories.

Around 22% of all vocalizations are annotated with two categories. We analyzed the most frequent co-occurrences of such multi-categorical annotation. Among

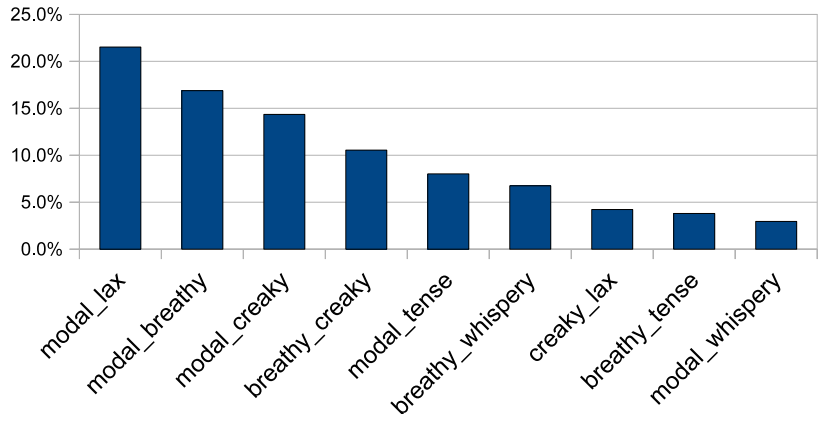


Figure 7.11: The distribution of co-occurrence of primary and secondary categories

vocalizations annotated with multi-categories, the most frequent co-occurences are *modal_lax* (21.5%), *modal_breathy* (16.9%), *modal_creaky* (14.3%) and *breathy_creaky* (10.5%) (see Figure 7.11).

Inter-rater agreement of voice quality annotation

In order to determine the reliability of voice quality annotation, a subset of 109 listener vocalizations is annotated by another student assistant with the same instructions. Cohen's kappa reliability of primary category annotated by both annotators is 0.55.

As we allowed for more than one category per instance, we computed Cohen's Kappa separately for each category, treating annotations as a binary "present/absent" feature. On this basis, we computed Kappa for each voice quality category. As shown in Table 7.6, the Kappa values for the most frequently used voice quality categories *modal*, *breathy*, *creaky*, *whispery*, *lax* and *tense* were 0.72, 0.76, 0.85, 0.94, 0.70 and 0.99 respectively.

7.7 Discussion

In this chapter, we described an exploratory procedure in order to annotate meaning and behavior. The procedure was helped us to come up with a set of suitable categories. However, the behavior annotation showed higher inter-rater agreement when compared to meaning annotation. Therefore, the procedure for the behavior annotation looks promising but not for the meaning annotation. One possible reason for the

voice quality	Cohen's Kappa
Modal	0.72
Breathy	0.76
Creaky	0.85
Whispery	0.94
Lax	0.70
Tense	0.99

Table 7.6: Cohen's Kappa reliability measures for voice quality annotation

less reliable meaning annotation could be the large set of meaning descriptors that we finally ended up with. Due to the *open-ended* annotation strategy, we finally used 37 organically grown large set of meaning descriptors. The possibilities to consolidate the list meaning descriptors have to be investigated.

The approach also seems to provide more understanding of mapping between meaning and behavior. As laughter is a simple description of behavior, we tried to find a correlation between meaning and laughter that we have obtained from the initial ABL annotation. Treating in the first instance laughter as a single behavioral category, we can investigate the meaning categories associated with it (Figure 7.12). It can be seen that laughter nearly exclusively occurs with amusement, and that much of it is friendly. However, some laughter is not friendly, and even scornful. For a synthesis system, it would be extremely important to know whether the laughter itself, in isolation, contains the "friendly" vs. "scornful" elements of meaning or if these have been derived from the context. If appropriate, then, several kinds of laughter should be distinguished in order to obtain as simple as possible a mapping between meaning and behavior.

Similarly, investigating the impact of several behavior properties of the vocalizations, such as *intonation* and *segmental form*, on their perceived meaning would be important; this may be helpful for the realization strategies to synthesize vocalizations. Due to the fact that the reliability of meaning annotation is low, we have not gone through detailed analysis of the relation between meaning and behavior in this chapter. Nevertheless, we made an attempt to investigate the relation between the meaning and behavior in Chapter 8.

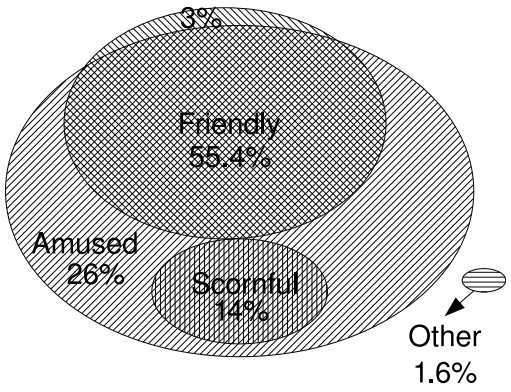


Figure 7.12: Distribution of meaning categories on laughter vocalizations.

7.8 Summary

In this chapter, we explored a methodology for annotating listener vocalizations. We started with a simple affect-backchannel-laughter annotation scheme. We then continued with an open annotation through informal descriptions of behavior and meaning. In the later stages, we used a joint categorical description for meaning and reference, in which the meaning is described using affective-epistemic categories, and reference annotation is based on Bühler's Organon model. In addition, we also explored methodologies to annotate behavioral properties such as intonation and voice quality. Clustering techniques were adapted to annotate intonation contours, while Laver (1991)'s voice quality categories were used for voice quality annotation. According to inter-rater agreements, behavior annotation seemed to be reliable when compared to meaning annotation.

The generally low inter-rater agreement shows that further work is needed before the meaning and reference annotation scheme can be considered a reliable tool for describing data. Improvements can be expected from a consolidation of the large set of meaning categories into a smaller set of clearly distinguishable categories, as well as improved annotation instructions.

Chapter 8

Multi-dimensional meaning annotation

The meaning annotation of listener vocalizations is a crucial step towards the synthesis of these vocalizations. The previous chapter has presented an open-ended exploratory study in order to identify the list of possible meanings available in a database of German listener vocalizations. Although such study helped us to identify a possible list of meanings available in the corpus, we made a few observations through the study. They are the following:

1. The agreement in meaning annotation provided by two raters on a small set of sub-corpus is very low, which indicates that a single rater annotation is undesirable. Another reason for the low agreement could be that the *open-endedness* of the task created confusion among the annotators.

2. Due to the open-ended annotation scheme, a large number (i.e. 37) of meaning categories were finally used; however, only a small set (i.e. 11) of the meaning descriptors were used on at least 5% of vocalizations available in the corpus. This indicates that the size of meaning descriptors list is redundant.

3. As the exploratory study used only a categorical annotation approach, it was not allowed to know vocalizations' *appropriateness* for each of the available meanings, which is a most desirable measure for synthesis of listener vocalizations.

In addition, several other studies were attempted to understand meanings of vocalizations as described in Chapter 2. However, none of them was focussed on *appropriateness* measures of meanings. An integrative account of all these studies must be considered in a bigger picture. It requires the following sequence of steps: (i) identification of suitable meaning descriptors; (ii) annotation of appropriateness for each meaning descriptor; (iii) identifying a typical impression of meanings for each vocalization; (iv) analyzing the impact of behavioral properties such as *segmental form* and *intonation* on perceived meaning. We attempt the above steps in this chapter.

In order to synthesize an appropriate listener vocalization, we require two kinds of information about each of the available vocalizations:

- A typical impression of the meaning that the vocalization could convey;
- How appropriate is the vocalization for a given meaning.

This chapter describes a systematic study of vocalizations' meanings. We propose a multi-dimensional annotation approach aimed at obtaining appropriateness ratings of each vocalization for each of the meanings. We conduct a listening test where multiple subjects annotate (characterize) a set of listener vocalizations using a multi-dimensional set of meaning descriptors. Typical impressions on context-independent meaning of listener vocalizations are being investigated. We also analyze the relevance of behavior properties for the meaning perception of listener vocalizations.

This chapter is organized as follows. In Section 8.1 the vocalizations database used in this study is described. Section 8.2 describes our approach for multi-dimensional meaning annotation, which includes the approach for stimuli selection and perception experiment. In Section 8.3 main results are discussed and in Section 8.4 findings are summarized.

8.1 Experimental corpus

Table 8.1 shows the database of vocalizations, which is recorded by four professional British actors, used for multi-dimensional meaning annotation. The complete details of the database are described in Chapter 6.

	Prudence	Poppy	Spike	Obadiah
Corpus duration (in minutes)	25	30	32	26
number of vocalizations	128	174	94	45

Table 8.1: British English listener vocalizations recorded for the four SAL characters

8.2 Approach

Annotation of a long list of meanings on scales (i.e. *appropriateness*) for all listener vocalizations is a tedious and time consuming process. Instead, annotation of selective vocalizations with a reduced set of meaning descriptors would be more cost effective. Therefore, we first propose an approach to consolidate the meaning descriptors by considering lessons learned from the exploratory study. Secondly, we describe a semi-automatic procedure used to select representative vocalizations in the corpus. Finally, we discuss a web-based perception experiment conducted for annotating *appropriateness* measures of the meanings. This section describes all these steps in detail.

8.2.1 Consolidating meaning descriptors

As described above, we intend to consolidate the list of meaning descriptors obtained in previous chapter (Chapter 7) in order to attain a balance between: the efforts needed for the annotation; and the percentage of vocalizations covered by the consolidated list.

We started by establishing a list of meaning dimensions, based on three sources: the most frequent categories in the exploratory annotation study on German listener vocalizations; the most frequently used annotations of the SEMAINE corpus (McKeown et al. 2010) – a large and annotated collection of dialogue of the SAL domain; and a set of affective-epistemic descriptors used to describe visual listener behavior (Bevacqua et al. 2007). The following are the meaning dimensions of these three sources:

- **Frequent categories used in exploratory annotation** (from Chapter 7) – surprised, anticipating, irritated, outraged, despondent, scornful, amused, tentative, doubtful, interested, thoughtful, compassionate, friendly

- **List of SEMAINE Categories** (McKeown et al. 2010) – expectation, intensity, anger, happiness, sadness, disgust, contempt, amusement, certain, agreeing, interested, at ease, thoughtful, concentrating, shows solidarity, shows antagonism and so on.

- **Categories used to describe visual listener behavior** (Bevacqua et al. 2007) – agree-disagree, interested-not interested, like-dislike, believe-disbelieve, accept-refuse, understand-don't understand

We also made sure that the consolidated list of categories is derived from three different backgrounds: they are emotional categories (Ekman 1999), Baron-Cohen's epistemic mental states (Baron-Cohen et al. 2004) and Bales Interaction Process Analysis (IPA) (Bales 1950). The reason is simple: the listener vocalizations, as described in Chapter 2, convey affective states, epistemic states and turn-taking cues; and they include cognitive, social and discourse regulatory functions. According to this author's knowledge, the three backgrounds are the best sources available to cover these states and functions. Emotional categories convey affective meanings; epistemic states can be used to represent attitudinal mental states of listener; and IPA labels can be used to identify social meanings in dialogue.

Descriptors	**Scale type**	**Source**
anger	unipolar	Emotional categories
sadness	unipolar	
amusement	unipolar	
happiness	unipolar	
contempt	unipolar	
solidarity	unipolar	IPA categories
antagonism	unipolar	
(un)certain	bipolar	Baron-Cohen's categories
(dis)agreeing	bipolar	
(un)interested	bipolar	
(high/low)anticipation	bipolar	

Table 8.2: Consolidated list of meaning descriptors used in this study

The three sources were consolidated into a list of 11 descriptors as shown in Table 8.2. The table shows the scale type (unipolar/bipolar) of meaning descriptors.

8.2.2 Stimuli selection

The stimuli are selected based on a semi-automatic clustering of intonation contours. For clustering vocalizations according to intonation, a contour was automatically computed for each vocalization by fitting a 3rd-order polynomial to f0 values extracted using the Snack pitch tracker (Sjölander 2006). Polynomials can approximate intonation contours of speech signal in unvoiced regions. Separately for each speaker, we used K-means clustering of intonation contours to identify the vocalizations with a similar intonation.

Two sets of stimuli were manually extracted from the clustered data for the purpose of selecting representative vocalizations that cover the maximum number of possible segmental forms and intonation contours. We aimed for two sets that contain, on one hand, stimuli with the same segmental form (as determined from the single-word description) varying in intonation (identified in the following as *fixed segmental form*); and on the other hand, stimuli with the same intonation (flat intonation contour) and varying in segmental form (henceforth, *fixed intonation contour*). Thus we manually selected samples from clusters as follows: (i) in order to get wide range of contour shapes, we selected one or two representative samples from each cluster with same segmental form (i.e. *yeah*); (ii) we selected samples with different segmental forms from a single cluster where contour shape is constant. Table 8.3 shows the number of selected stimuli for the experiment.

Character	*Fixed segmental form*	*Fixed intonation contour*
Poppy	15	8
Spike	10	9
Obadiah	5	8
Prudence	8	9
Total	38	34

Table 8.3: Character wise number of vocalizations selected for meaning annotation

8.2.3 Perception experiment

Scale-based ratings capture inherent ambiguity more than forced-choice test. We designed a web-based perception study for participants (see Appendix A). The first page provided instructions, the second page collected demographic information and the following pages present the audio and rating scales one at a time, as shown in Figure 8.1. The stimuli were presented to the participants in a random order for eliminating order and fatigue effects. Participants could play the audio as many times as they liked before providing meaning ratings. A 5-points Likert scale for each meaning was used: from 1 (absolutely no attribution) to 5 (extremely high attribution) for unipolar meaning categories; from -2 (extremely negative attribution) to +2 (extremely positive attribution) for bipolar meaning categories. "No Real Impression" option was provided for each meaning scale in case the participant is unsure.

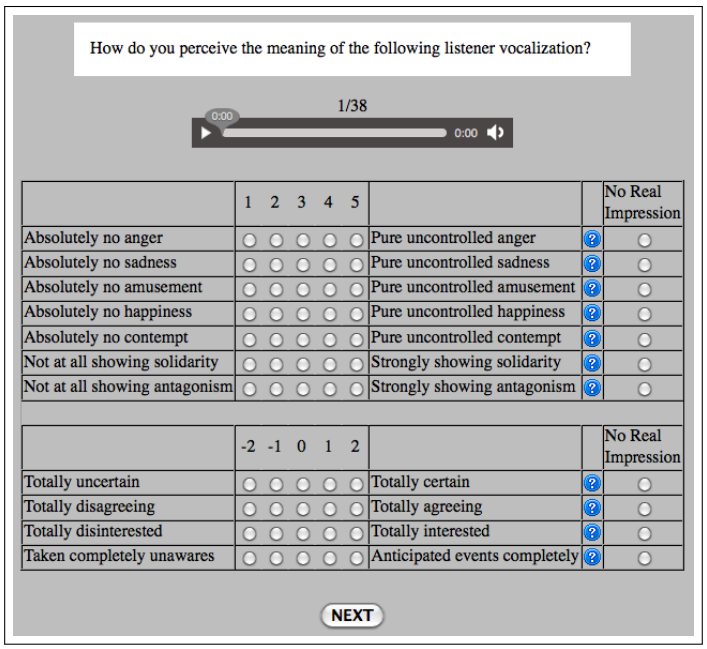


Figure 8.1: A screenshot of the web page for the perception study

44 participants (20 women, 24 men) took part in the annotation study. 22 participants provided ratings for the vocalizations in test set *fixed segmental form* (9 women, 13 men) and 22 participants rated vocalizations in test set *fixed intonation contour* (11 women, 11 men). The ratings obtained for each of the stimuli is summarized in Appendix B.

8.3 Results and discussion

In order to study each of the vocalizations per meaning, we first introduce the term *meaning-vocalization* combination that is used in the rest of this chapter. Each vocalization can convey maximally 11 meanings used in the corpus annotation. One stimulus indicates 11 *meaning-vocalization* combinations. For example, in the case of Prudence (see Table 8.4), 187 *meaning-vocalization* combinations (17 stimuli * 11 meaning categories) were available for analysis. Such tables for all four characters can be found in Appendix C.

8.3.1 High versus Low agreement

Table 8.4 shows the high variability on agreement of *meaning-vocalization* combinations for Prudence. In this table high agreement is identified with circles or arrows and low agreement is identified with a dot (·). In order to identify high agreement versus low agreement of *meaning-vocalization* combinations, we computed the interquartile range (IQR) of ratings provided for each combination. We considered that a combination has high agreement if the IQR of the combination is less than one third of the meaning scale range. In other words, a combination has high agreement if more than 50% of the raters agree within one third of the meaning scale range. The high agreement combinations indicates typical impression of the meaning on the vocalization.

Table 8.4 shows that the number of low agreement annotations (identified as ·) are higher in the *fixed intonation contour* set when compared to the *fixed segmental form* set for Prudence. The same tendency was observed when taking into account all the vocalizations in our corpus, that is 792 (72 stimuli * 11 categories) *meaning-vocalization* combinations, from which 418 combinations belong to the *fixed segmental form* set and 374 belong to the *fixed intonation contour* set. Figure 8.2 shows a global picture of high agreement versus low agreement combinations for all the corpus. While around 60% of the *fixed segmental form* combinations show high agreement, only 40% of the *fixed intonation contour* combinations show high agreement. This seems to indicate that the participants perceived more distinguishable information from intonation when compared to segmental form. In other words, this evidence indicates that the intonation contour is highly relevant for signaling meaning when compared to phonetic segmental form.

Fixed segmental form												
segmental form	intonation-contour	voicequality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	⁀	modal	○	·	○	○	·	↑	○	↑	↑	○	○
yeah	⁀	modal	○	⇑	○	○	○	·	○	○	○	·	○
yeah	‾	creaky	○	·	○	○	·	·	○	↑	↑	○	↑
yeah	⁀	modal	○	○	·	·	○	⇑	○	↑	↑	↑	·
yeah	‾	modal	○	·	○	○	○	·	·	○	↑	·	·
yeah	‾	modal	○	○	⇑	⇑	○	⇑	○	↑	↑	↑	·
yeah	‾	creaky	○	·	○	○	○	·	○	○	○	↓	○
yeah	‾	modal	○	·	○	○	·	↑	·	·	·	↓	○

Table 8.4: Fixed segmental form set: Segmental form, intonation contour and meaning of Prudence's stimuli. Meaning-vocalization combination is represented using the following symbols.
○ : vocalization is not appropriate for the meaning;
↑ or ↓ : vocalization is somewhat appropriate;
⇑ or ⇓ : vocalization is very appropriate for the meaning;
· : the annotation has low agreement (we can not conclude on appropriateness);
↓ and ⇓ : negative sides of bipolar scales

Fixed intonation											
segmental form intonation-contour voicequality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
tsyes ‿ modal	·	·	·	·	·	·	·	↑	○	·	·
tsyeah ‿ modal	·	·	·	○	·	·	·	·	·	·	○
mhm ‿ modal	·	·	○	○	·	↑	·	·	○	·	·
yeah ‿ modal	·	·	○	○	·	·	·	↑	·	·	·
yes ‿ modal	·	·	○	○	·	·	·	○	○	○	↑
right ‿ modal	·	○	○	○	·	·	·	·	·	·	·
tsright ‿ modal	·	·	○	○	·	·	·	↑	·	·	○
aha ‿ modal	○	○	·	·	·	⇑	○	↑	↑	↑	↑
tsgosh ‿ modal	○	○	·	·	·	·	○	○	○	○	○

Table 8.5: Fixed intonation contour set: Segmental form, intonation contour and meaning of Prudence's stimuli. Meaning-vocalization combination is represented using the following symbols.
○: vocalization is not appropriate for the meaning;
↑ or ↓ : vocalization is somewhat appropriate;
⇑ or ⇓ : vocalization is very appropriate for the meaning;
·: the annotation has low agreement (we can not conclude on appropriateness);
↓ and ⇓ : negative sides of bipolar scales

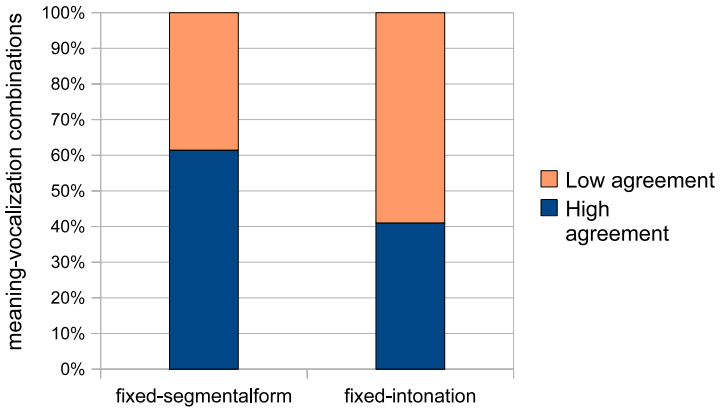


Figure 8.2: Percentage of high vs. low agreement *meaning-vocalization* combinations

Figure 8.3 shows high agreement versus low agreement combinations per meaning scale. In the case of *fixed segmental form*, the participants showed high agreement on *amusement, happiness, anger, certain* and *agreeing* meaning scales and low agreement on *contempt, solidarity and antagonism*. In case of *fixed intonation contour*, the participants did not show high agreement on any meaning scale, but they agree more on

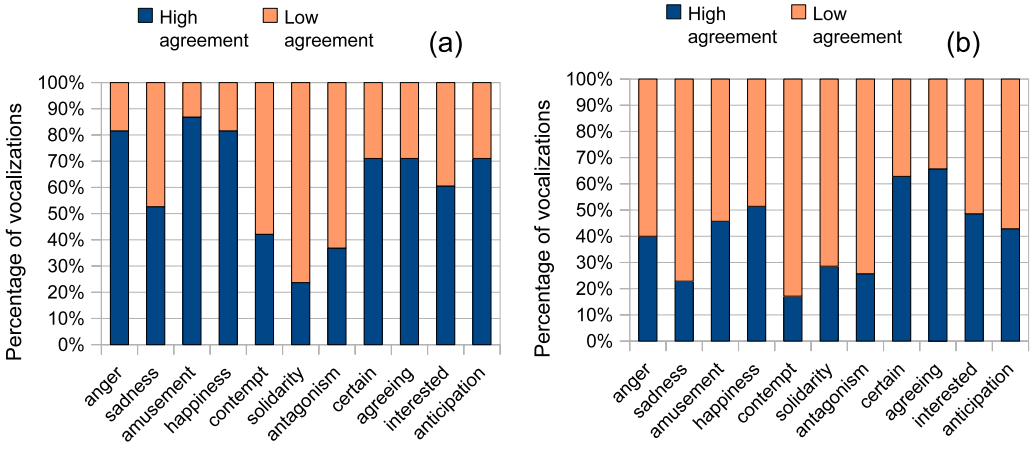


Figure 8.3: Percentage of high vs. low agreement vocalizations per each of the meaning descriptors: (a) fixed segmental form (b) fixed intonation contour

certain and *agreeing* when compared to the other scales. The low agreement might be due to the inherent ambiguity of the listener vocalizations or due to the meaning descriptors.

The results seem to indicate that the affective meanings of listener vocalizations can be more distinguishable by intonation contour than by phonetic segmental form, since the emotional categories (anger, sadness, amusement, happiness, and contempt) showed higher agreement on the fixed segmental form set than on the fixed intonation contour set.

8.3.2 Appropriateness of high agreement annotations

Not all vocalizations with high agreement may be suitable to convey a specific meaning for synthesis. In this work the suitability of a *meaning-vocalization* combination is calculated by computing the median of ratings provided for that combination. In other words, we consider that a vocalization is very appropriate for a specific meaning if the median of ratings is closer to the positive end of the scale. However, we can not conclude about suitability of low agreement ratings.

We distinguish three levels of appropriateness based on where the participants tend to agree on the meaning scale. A *meaning-vocalization* combination is very appropriate if the participants tend to agree on positive (in case of unipolar and bipolar scales) or negative (in case of bipolar scale) end of meaning scale. The combination is not appropriate to convey the meaning if they tend to agree on '0'. In other words, we can say that the combinations are "very appropriate", "somewhat appropriate", and "not appropriate" when the median is greater than two third of meaning scale, between one third and two third, and less than one third respectively. Among high agreement *meaning-vocalization* combinations available in our corpus, it was found that, 7.2% (30) are very appropriate, 22.4% (93) are somewhat appropriate, and 70.4% (293) are not appropriate combinations. This result is highly relevant in speech synthesis, that is, one vocalization can be "not appropriate", "somewhat appropriate" or "very appropriate" for several different meanings at the same time. These three categories can be used, for example, in an algorithm for unit-selection synthesis (i.e. vocalization selection) that considers appropriateness to realize a particular intended (target) meaning. Such a unit-selection algorithm is presented in Chapter 9.

8.3.3 Inherent ambiguity of listener vocalizations

According to Table 8.4, the vocalization *aha* can convey 5 meanings (*solidarity, certain, agreeing, interested, anticipation*), whereas the vocalization *right* does not convey any meaning available in our descriptors. Figure 8.4 shows the histogram of possible meanings for the listener vocalizations in our corpus. Among 72 stimuli, 14 vocalizations (19.5%) convey no meaning, 27 (37.5%) convey a single meaning, and the remaining 31 (43%) convey multiple meanings. On average, a single vocalization in this corpus conveys 1.68 meanings, which confirms the multifunctional nature noted by Schegloff (1982a) and McCarthy (2003). Indeed the inherent ambiguity of listener vocalizations is a very interesting feature to explore in speech synthesis, because a single vocalization can be used in multiple instances.

More than 80% of the vocalizations in our corpus can convey at least one meaning among the consolidated list. However, around 20% of the vocalizations appear that they are not able convey any meaning. So this might indicates that the consolidated list of meaning descriptors may not be sufficient to represent all of the vocalizations.

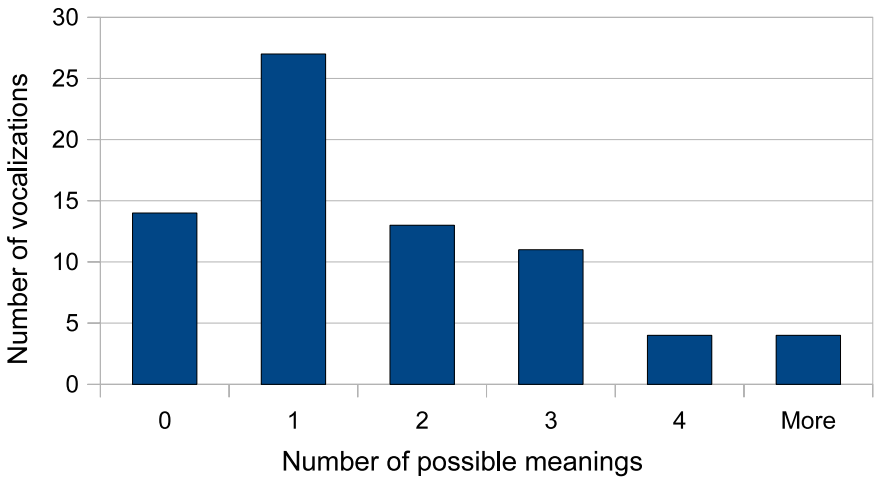


Figure 8.4: Histogram of multiple meanings

8.4 Summary

In this chapter, we explored a multi-dimensional annotation methodology to annotate listener vocalizations in view of conversational speech synthesis. We conclude the following issues from this study: (i) this methodology can provide a typical impression of meanings from high agreement annotations; (ii) unit-selection algorithms can benefit from the annotation of meaning on scales: it captures appropriateness of listener vocalizations for a given meaning; (iii) one vocalization can convey multiple meanings, which is useful for the usage of the same vocalization in several instances; (iv) the evidence seems to indicate that the intonation contour is highly relevant for signaling meaning when compared to the phonetic segmental form - in support for improving acoustic variability using imposed-intonation contours.

Chapter 9

Realization

The realization strategy is crucial in corpus-driven synthesis of listener vocalizations. *Appropriateness* and *naturalness* would be key words for such strategies. In other words, the synthesized speech is expected to be not only appropriate for a given request but also of good quality. This chapter is aimed to come up with a good realization approach to synthesize listener vocalizations. To begin, we summarize what we already have by now as follows:

- Natural and spontaneous listener vocalizations
- Annotation of symbolic features for each of the vocalizations
  - Meaning appropriateness measures
  - Segmental form & (quasi-)phonetic alignment
  - Intonation
  - Voice quality

We have vocalizations and their symbolic features, as listed above. Briefly, the task of realization is to generate appropriate vocalizations when a user requests them with symbolic features. In order to standardize the user requests, Section 9.1 proposes a markup specification for MARYXML. Section 9.2 describes a simple unit selection algorithm to select the most suitable one from the available vocalizations for a given request; it also discusses the drawbacks of this approach and suggests ideas for possible improvements. We then propose an enhanced version of unit selection approach which

uses signal processing techniques to overcome the drawbacks, see Section 9.3. We explain the efforts in technical aspects to enable MARY TTS to synthesize listener vocalizations in Section 9.4. Finally, Section 9.5 summarizes this chapter.

9.1 Markup specification

To support the generation of non-verbal and quasi non-verbal vocalizations such as backchannels, a new element <vocalization> is introduced into the MARY-specific markup format MaryXML (Pammi et al. 2010). It allows a user to request a vocalization based on the following criteria:

- meaning: the intended meaning of the vocalization;
- intonation: the type of intonation contour used on the vocalization;
- voice quality: the voice quality used with the vocalization;
- name: a description of the segmental form of the vocalization.

An example of the markup request is shown in Figure 9.1. All of the attributes of the <vocalization> tag are optional; if an attribute is not given, this means that the search is not constrained on that level.

```
<maryxml>
  <voice name="dfki-poppy">
    <vocalization
      name="yeah"
      meaning="agreeing"
      intonation="rising"
      voicequality="modal"/>
  </voice>
</maryxml>
```

Figure 9.1: Example of MaryXML markup requesting the generation of a vocalization.

Table 9.1 lists the possible values currently supported for each of the criteria; note that not all values are available with every voice, except *name* attribute which depends

on the segmental forms available in the corresponding voice. Around 17 discrete meaning values are used as markup specifications for runtime synthesis. These meaning attributes are extracted from 11 meaning dimensions used in annotation (i.e. Chapter 8). The extreme (completely positive or completely negative) ends of each of the scales are represented with one of the discrete meaning descriptors. The advantages of discrete categories being used by markup requests are: (i) the discrete meaning categories can easily be used by the user (i.e. markup) requests; (ii) we can easily find appropriate candidates for a given target, because the target have a single meaning in order to compare several candidates that have different appropriateness ratings for that meaning.

The list of values for intonation and voicequality result from a provisional annotation (as described in Chapter 7) of the vocalizations of a German professional actor. Clearly, the list of these values will need to be broadened in the future; notably, the values for describing intonation contours are insufficient. However, given the fact that linguistically motivated descriptions of intonation, such as ToBI (Silverman et al. 1992), are probably inadequate for the emotional and discourse-oriented meanings found in listener vocalisations, it is not straightforward to select an appropriate descriptive scheme for intonation contours. Nevertheless, as it stands, the list of values provides a reasonable standing point for developing the synthesis which is the core topic of the synthesis.

9.2 Simple unit-selection algorithm

In this section, we describe a simple vocalization (in the following chapter, we also call them *units*) selection algorithm, which is a simple playback approach that works on top of a cost function based selection criterion. As described earlier, the notion of *target* is central in unit selection algorithms. The main objective of this algorithm is to identify a best matching candidate among all candidate units for the intended target. As seen in Figure 9.2, the workflow of the algorithm can be described in four steps: (i) preparation of candidate units; (ii) preparation of target unit; (iii) applying a cost function; (iv) realization of the lowest cost unit. The further section describes these steps in detail.

Attribute	Possible values
meaning	anger, sadness, amusement, happiness, contempt, certain, uncertain, agreeing, disagreeing, interested, uninterested, low-anticipation, high-anticipation, low-solidarity, high-solidarity, low-antagonism, high-antagonism
intonation	rising, falling, high, low
voicequality	modal, creaky, whispery, breathy, tense, lax
name	`Depends on recordings of the voice: tokens like` yeah, yes, mhmh, mhm, right, tsright, tsyeah, aha etc.

Table 9.1: Values currently supported for each of the attributes of the `<vocalization>` element in MaryXML.

STEP 1: preparation of candidate units

In this algorithm, the candidate units represent all of the available vocalizations and their annotations in the database. Each candidate unit represents one recorded vocalization and its following annotation properties (as shown in Figure 9.2):

- Appropriateness of meaning: median value of all subjective scale-based ratings
  - unipolar (within the range from 0 to 5) for the descriptors *anger, sadness, amusement, happiness, contempt, solidarity*, and *antagonism*.
  - bipolar (within the range from -2 to 2) for the descriptors *certain, agreeing, interested*, and *anticipation*.
- Segmental form (i.e. one-word description) of the vocalization
- Intonation of the vocalization
- Voice quality of the vocalization

STEP 2: preparation of target unit

In this step, a target unit that represents the ideal vocalization for a given markup request is constructed. A target unit is created from the markup request, containing

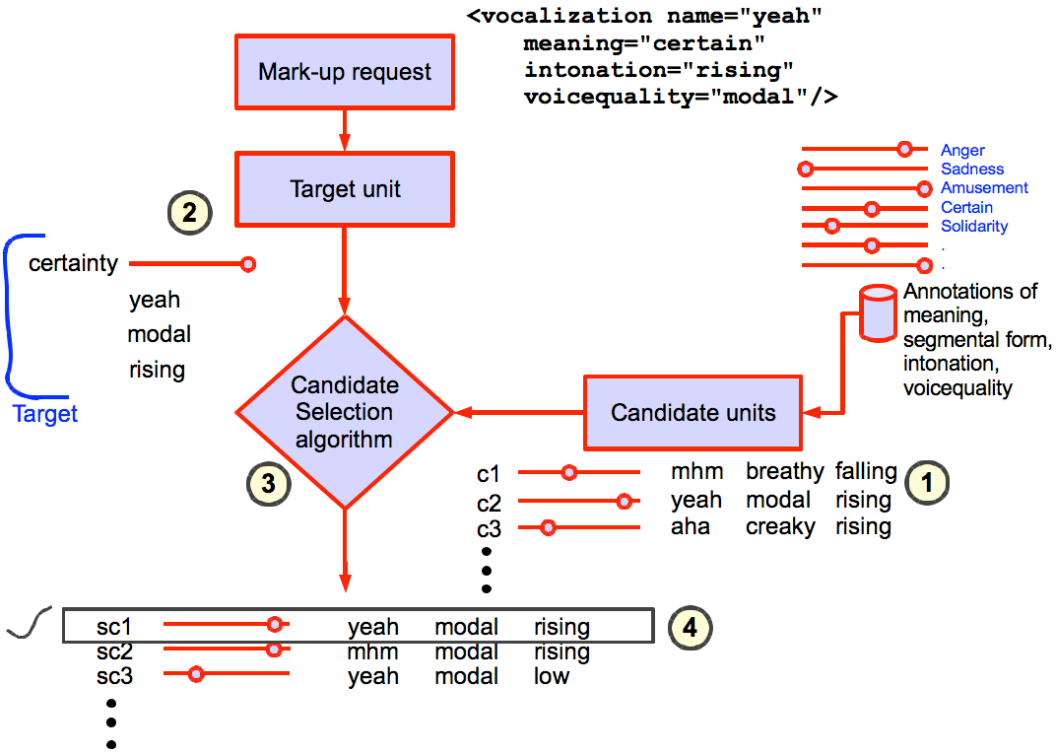


Figure 9.2: A simple unit selection strategy to synthesize listener vocalizations

as features the values given in the markup attributes, or "unspecified" if the respective attribute is omitted.

Consider the following markup request as an example:

```
<vocalization name="yeah" meaning="certain"
intonation="rising" voicequality="modal"/>
```

The idealistic situation for the above mark-up request is the synthesis of *yeah* vocalization; which has high appropriateness for the meaning *certainty*; which has *rising* intonation contour; and which has *modal* voice quality.

STEP 3: cost function based selection of candidates

Unit selection principles are used to select the best candidate vocalization for a given request. A unit in this case represents the entire vocalization; therefore, our cost function uses only target costs, no join costs.

The target cost is a weighted sum of feature costs. All candidate units are sorted according to the suitability with the target unit, where the suitability is computed based on a target cost function. In other words, the lower the target cost, the better the candidate matches the requested target. The cost function can be written mathematically as follows:

$$C(u_i) =< w, c(u_i) > \tag{9.1}$$

where u_i is the candidate unit i; c is the cost vector containing several feature costs; and w is the weight vector for the features.

The elaborated mathematical equation for the available features can be written as the following:

$$C(u_i) = W^{\mathrm{T}} * \begin{pmatrix} segCost(u_i) \\ f0Cost(u_i) \\ vqCost(u_i) \\ meaningCost(u_i) \end{pmatrix} \tag{9.2}$$

where $segCost$ is segmental form cost; $f0Cost$ is intonation cost; $vqCost$ is voice quality cost; $meaningCost$ is meaning cost; and W is the corresponding weight vector.

STEP 4: realization of the lowest cost unit

A candidate unit with the lowest target cost is synthesized in the final step. The realization is a simple 'play-back' approach, which means that the entire recorded vocalization is realized as synthetic speech. As a result, the naturalness of synthetic speech in this approach is the same as for human speech.

9.2.1 Drawbacks

Problem with unseen data

An important limitation with the simple unit selection approach to the synthesis of listener vocalizations is the fact that we can only generate the vocalizations that have been recorded. If we require additional vocalizations, such as an existing segmental form but with a meaning that had not been produced by the original speaker during the

recording session, then this algorithm can only produce the vocalization most similar to the target – which may not be acceptable. In the example shown in Figure 9.3, if we assume that the first candidate unit is not available in the recorded corpus, the algorithm realizes the token *mhm* instead of the intended segmental form *yeah*, which is not desirable. Alternately, the third candidate could have been chosen by a different cost function, which would yield the target segmental form but a meaning that differs markedly from the requested one.

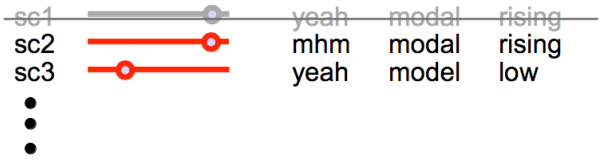


Figure 9.3: An example case for unseen data

Limited acoustic variability

This algorithm is being suffered with limited acoustic variability due to the above mentioned 'unseen data' problem. As described in Chapter 6, collecting data with much acoustic variety is a difficult task. We may not be able to record vocalizations that cover all dimensions of segmental form and meaning (i.e. with all possible tokens with all meanings). The realization algorithm should be able to somehow handle this issue, however, the simple selection algorithm can not improve the acoustic variability. If the recorded material does not have much acoustic variety in the corpus, this algorithm performs not good enough. For example, if we request the token *yeah* with high appropriateness for the meaning *uncertainty*, and if it is not pre-recorded, the algorithm can not do much to synthesize it. The algorithm can realize either the token *yeah* with other meaning or other token with the meaning *uncertainty*, but not both. The synthesis of the intended token with the intended meaning is not always achievable.

9.2.2 Ideas for improvement

In order to extend the space of options, we propose a methodology to make use of signal modification techniques. The idea behind the methodology is explained with

the example shown in Figure 9.4. According to the example, the target was the token *yeah* with the meaning *certainty*. The best possible candidates are (i) the token *mhm* with high *certainty*; (ii) the token *yeah* with low *certainty*. Each candidate thus has only part of the desired properties. The proposed idea is now to extract the behavior properties of the token *mhm*, which is closer to the intended meaning, and impose them onto the token *yeah* in order improve its appropriateness for the meaning *certainty*. In generic terms, the methodology is to extract behavior properties such as *intonation contour* from most appropriate vocalization for the intended meaning, and to impose these properties onto the most appropriate token for the target. This idea is developed into a new realization strategy in the next section.

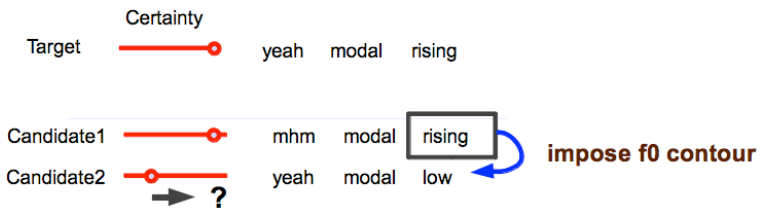


Figure 9.4: An idea for improvement in unit selection

It is a well known fact that signal modification degrades the naturalness of the speech signal. The more signal modification leads to the least naturalness of synthesized speech. A secondary constraint in developing the realization strategy is therefore the challenge to find a tradeoff between modification of signal and naturalness of synthesized speech.

9.3 Unit-selection algorithm: new method

The objective of the synthesis algorithm is to synthesize the requested segmental form with the intended meaning requested by MARY XML even in cases where the requested combination is not available from the recordings. This section describes an extended algorithm for selecting both candidate units and intonation contours separately, and for combining them using signal modification techniques.

The basic idea of the new approach, as shown in Figure 9.5, is to combine unit selection principles with signal post-processing to impose a suitable intonation contour

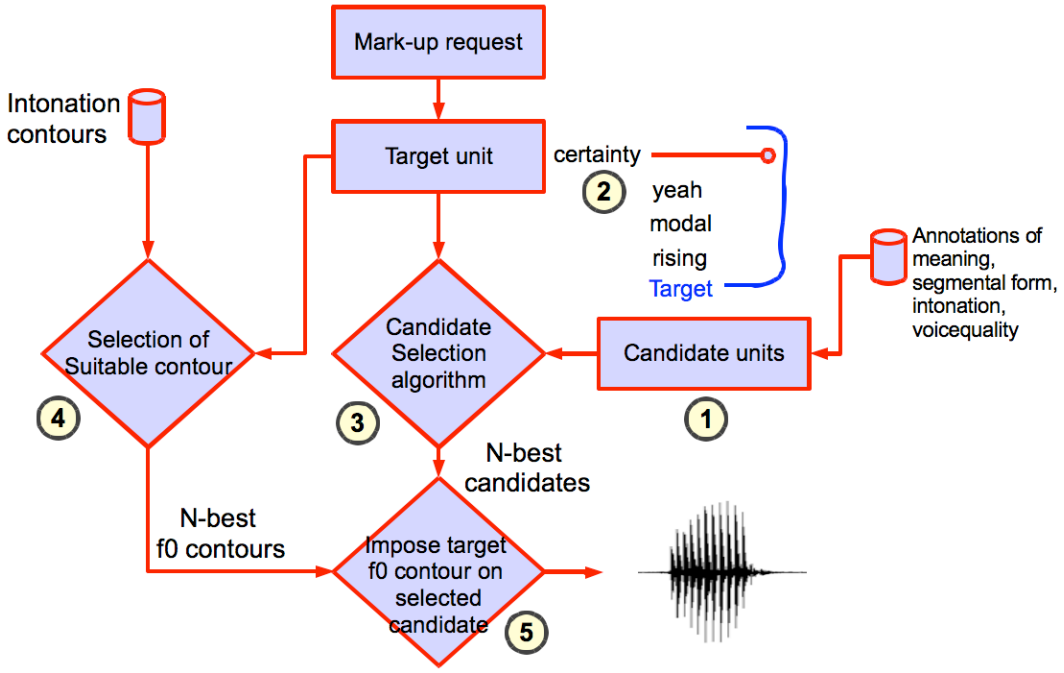


Figure 9.5: A new method for synthesis of vocalizations using imposed intonation contours

onto an approximately suitable vocalization. Given a request formulated using speech synthesis markup, we construct a target unit representing the ideal vocalization. A target cost function is used to select the best candidate from among the available recordings in the given voice. The target unit is also used to select a suitable intonation contour, which is then imposed onto the selected unit.

As shown in Figure 9.5, the new method consists of five steps. The first two steps (**STEP 1** and **STEP 2**) are exactly the same as in the simple unit selection approach. The remaining steps are discussed in the subsequent sections in detail.

STEP 3: Selecting the best unit candidates

This step is somewhat different from the corresponding step in the simple selection algorithm. As said earlier, the target cost is a weighted sum of feature costs. In case of unit candidate selection, we give more weight to the segmental form rather than the meaning of the vocalization. This reflects the choice that, if at all possible, the

segmental form of the selected vocalization should be realized as requested.

The cost function uses a manually created similarity matrix for each feature. Compared to the classical evaluation function, which assigns cost 0 for equal values and cost 1 when values are different, the similarity matrix has the advantage that it can capture the degree of similarity between feature values. Where a unit exactly matching the target is not available, it is preferable (i.e. less costly) to use a similar unit rather than a very different one. For example, the similarity between the segmental forms 'yeah' and 'myeah' is high (resulting in low cost), whereas the similarity between 'yes' and 'no' is low, and thus results in high cost for that feature. We manually fill the similarity matrices and assign the weights to the different features. The special value "unspecified" has cost 0 for all feature values.

$$UC(u_i) = W_{uc}^{\mathrm{T}} * \begin{pmatrix} segCost(u_i) \\ f0Cost(u_i) \\ vqCost(u_i) \\ meaningCost(u_i) \end{pmatrix} \tag{9.3}$$

where UC(u_i) is the cost to select unit candidate u_i; and W_{uc} is the weight function for candidate selection.

The candidate selection algorithm selects a configurable number of best candidates.

STEP 4: Selecting the best contour candidates

The contour selection algorithm selects a number of best contour candidates among the available vocalizations using the same algorithm as for the unit candidates, except that different weights are used. Whereas the unit candidate selection gives more weight to the segmental form, the contour candidate selection gives more weight to the meaning features and zero weight to the segmental form and voice quality. Therefore, contour candidates are selected irrespective of their segmental forms.

$$CC(c_k) = W_{cc}^{\mathrm{T}} * \begin{pmatrix} segCost(c_k) \\ f0Cost(c_k) \\ vqCost(c_k) \\ meaningCost(c_k) \end{pmatrix} \tag{9.4}$$

where CC(c_k) is the cost to select contour candidate c_k; and W_{cc} is the weight function for contour selection.

In order to be included in the list of contour candidates, an intonation contour should bring the combined vocalization closer to the target than any of the candidate units by themselves, i.e. it should actually reduce the resulting cost. In order to decide this, we compute the smallest *contour* cost for the *unit* candidates, CC_{min}; only intonation contours that have a cost not greater than CC_{min} are considered.

If there are no suitable contour candidates, the best unit candidate is synthesized without imposing an intonation contour.

STEP 5a: Selecting the best unit-contour pair

In this step, it is decided which of the contour candidates to impose on one of the unit candidates. Given a unit candidate u_i with original contour c_i and a contour candidate c_k, we define the Unit-Contour Cost UCC as the weighted sum of a merged target cost MC, reflecting an estimate of the similarity of the merged vocalization to the target, and an intonation cost IC, with weight factor α, which attempts to reflect the distortions caused by imposing the intonation contour:

$$UCC(u_i, c_k) = MC(u_i, c_k) + \alpha\, IC(c_i, c_k) \tag{9.5}$$

$IC(c_i, c_k)$ is computed as a distance between third-order polynomial approximations of the respective contours (Fujii, Kashioka, and Campbell 2003).

We compute the merged target cost using the formula,

$$MC(u_i, c_k) = W_{mc}^{\mathrm{T}} \begin{pmatrix} segCost(u_i) \\ 0 \\ vqCost(u_i) \\ \frac{1}{2} meaningCost(u_i) \end{pmatrix} + W_{mc}^{\mathrm{T}} \begin{pmatrix} 0 \\ f0Cost(c_k) \\ 0 \\ \frac{1}{2} meaningCost(c_k) \end{pmatrix} \tag{9.6}$$

where $segCost$ is segmental form cost; $f0Cost$ is intonation cost; $vqCost$ is voice quality cost; $meaningCost$ is meaning cost; W_{mc} is a weight function for merged cost calculation.

If all weights are configured to unity, Equation 9.6 becomes the following:

$$MC(u_i, c_k) = segCost(u_i) + f0Cost(c_k) + vqCost(u_i) + \frac{1}{2}(meaningCost(u_i) + meaningCost(c_k))$$

The pair of unit and contour candidates that minimizes the Unit-Contour Cost is selected.

STEP 5b: Imposing a target intonation contour

The selected unit and contour can be combined using signal modification techniques such as the Frequency-Domain Pitch-Synchronous Overlap Add (FD-PSOLA) algorithm (Moulines and Verhelst 1995), Mel-Generalised Log Spectral Approximation (MLSA) Vocoding, and Harmonics Plus Noise Model (HNM) vocoding. In order to make the synthesized vocalization insensitive to unvoiced regions and large pitch excursions, a third order polynomial approximation of the source contour is used as the target contour for imposition. A reduced copy of the original intonation is used in case of extreme pitch ranges to reduce the effect of distortions.

9.4 Enabling MARY to synthesize vocalizations

This section describes our efforts to implement the proposed algorithm in MARY framework. We first provide background information on three state-of-the-art signal modification techniques available in the MARY framework. We then discuss our implementations that are used to realize vocalizations using MARY framework. the use of these prosody modification techniques to impose the target intonation contours onto the vocalizations, according to the proposed realization strategy.

9.4.1 Signal modification techniques

This thesis uses the following prosody modification techniques that are already implemented in MARY framework. This section provides more information on them.

1. Mel-Generalised Log Spectral Approximation (MLSA or MGLSA) vocoding

2. Frequency domain pitch synchronous overlap-add (FD-PSOLA)

3. Harmonics Plus Noise Model (HNM) vocoding

MLSA vocoding

The MLSA or MGLSA (Mel-Generalised Log Spectral Approximation) vocoder is a digital filter for speech synthesis included in the HTS HMM-based synthesis engine (Tokuda et al. 2010). As described in Chapter 4, the HTS engine has been ported to Java based MARY framework, and the MLSA vocoder has been enhanced to use mixed excitation as in (Yoshimura et al. 2001b). The mel-generalised cepstral coefficients used in this vocoder are extracted with SPTK (Kobayashi et al. 2009) and the pitch contour with Snack (Sjölander 2006); pitch modification for the different vocalizations is realized resizing the target prosody to a candidate number of frames. Mixed excitation is realized with ten Fourier magnitudes for pulse excitation generation and five bandpass voicing strengths for better pulse/noise spectral shaping. Fourier magnitudes are calculated on the residual signal, obtained by inverse filtering, by detecting the first ten pitch harmonic peaks in the residual spectrum. Bandpass voicing strengths are estimated by filtering the signal into five frequency bands and calculating peak normalized cross correlation in each band. Voicing strengths and Fourier magnitudes were calculated with SPTK and Snack. Mixed excitation is calculated as follows: a pulse train is generated by inverse Fourier transform of the Fourier magnitudes for one pitch period. The pulse train and noise are passed through the five spectral shaping filters and then added together to give a full band excitation. For each frame, the frequency shaping filter coefficients are generated by a weighted sum of fixed bandpass filters. The pulse filter is calculated as the sum of each of the bandpass filters weighted by the voicing strength in that band. The noise filter is generated by a similar weighted sum, with weights set to keep the total pulse and noise power constant in each frequency band (McCree and Barnwell 1995).

Frequency domain pitch synchronous overlap-add

FD-PSOLA employs linear prediction to compute the spectral envelope and the excitation spectrum using pitch synchronous speech frames (Moulines and Verhelst 1995). Pitch modification is achieved by linear interpolation of the spectral envelope. The

residual spectrum is either shortened or expanded to match the new size of the spectral envelope. The modified spectral envelope and residual spectrum is then multiplied and the time-domain waveform is obtained by an inverse Fourier transform.

The major advantage of FD-PSOLA is its ease of implementation. Frequency domain operation makes it straightforward to perform spectral envelope modifications such as speaker identity transformation or normalization. On the other hand, FD-PSOLA lacks the functionality to provide explicit control of phase continuity. Therefore, when used in the context of concatenative synthesis, it may lead to discontinuities at concatenation boundaries.

HNM vocoding

The harmonics plus noise framework provides better control over phase continuity as described in (Stylianou 1996). HNM models the lower frequency portion of the speech signal using a set of harmonically related sinusoids. The difference between the original signal and the signal re-synthesized from the harmonic part is modeled as bandpass filtered random noise. The frequency boundary between the two bands is dynamically computed by analyzing and separating harmonic peaks from noisy peaks and then smoothing the result over consecutive speech frames.

Pitch modification is performed by computing a new set of harmonics according to the pitch scaling ratio while preserving the spectral envelope shape. The modified speech signal is obtained by interpolating phases and amplitudes across successive synthesis frames. Explicit phase interpolation reduces discontinuities at concatenation boundaries. As a variation of the original algorithm, the MARY framework uses the waveform corresponding to the noise part instead of employing the bandpass filtered noise model. This approach enables perfect reconstruction when no pitch modification is performed. The modified noise part generation uses simple overlap-add since no pitch modification is required for the noise part.

9.4.2 Realization with MARY framework

The workflow of realization of vocalizations can be divided into two stages: (i) Offline data processing; (ii) Runtime synthesis. This section describes the two stages.

Offline data processing

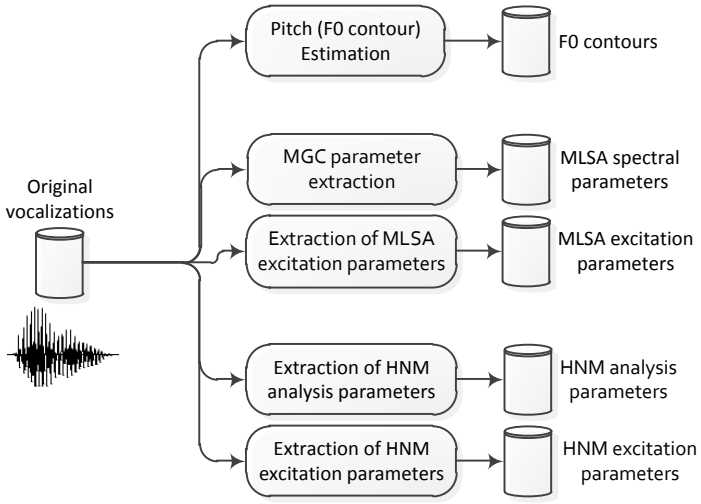


Figure 9.6: Parameter extraction from original vocalizations

The offline data processing is similar to voice creation process in TTS. In this stage, we analyze original vocalizations and prepare necessary files required at runtime speech synthesis. We developed several voice import components that are useful in building support for realization of listener vocalizations. Figure 9.6 describes processes of different such components; and it shows different stages of offline processing of original recordings. Initially, Praat or Snack is used to estimate intonation contours for the vocalizations, and third order polynomials are fitted on the contours in order to better handle unvoiced regions. Secondly, we extract MGC spectral parameters and MLSA excitation parameters. Finally, HNM analysis and excitation parameters are extracted to make use of them at runtime synthesis.

The voice import components that are responsible to extract these parameters also write them into a datagram file in an efficient timeline fashion in order to quickly extract the parameters with the index number of a vocalization.

Runtime synthesis

The main approach for the realization of listener vocalizations is described in Section 9.3. As described earlier, the approach consists two stages: selection of best candidate-contour pair (CCP); and the transportation of F0 contour of contour candidate (CC)

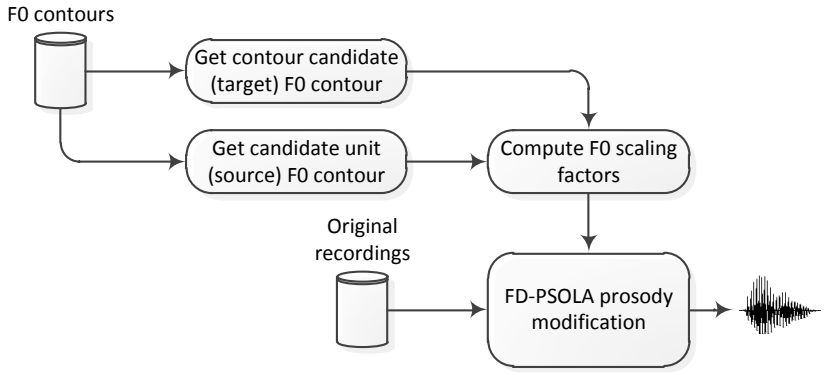


Figure 9.7: FD-PSOLA based prosody modification to synthesize vocalizations

onto the candidate unit (CU). This section discuss the practical implementation of the signal modification step.

The CCP consists of the index number of the CU and the index number of the CC. In order to transport F0 contour of the CC onto the CU, we use one of the three signal modification techniques explained in previous section. They are: (i) FD-PSOLA (ii) MLSA Vocoder (iii) HNM vocoder.

As shown in Figure 9.7, FD-PSOLA technique computes scale factors to change the CU intonation contour to the CC intonation contour. The prosody modification is realized by FD-PSOLA prosody modifier according to the scale factors.

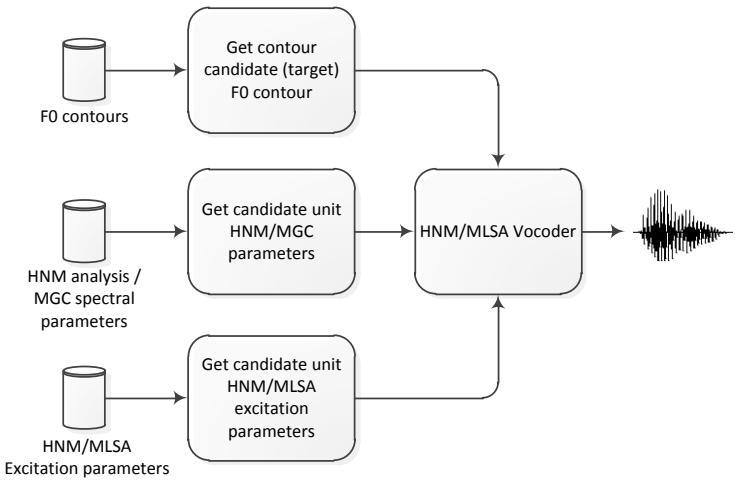


Figure 9.8: Vocoding strategies to modify prosody of vocalizations

Figure 9.8 shows the two vocoding procedures, MLSA based and HNM based methods, to modify the prosody of speech signals. The warped F0 contour of the CC, the spectral and excitation parameters of the CU are combined by MLSA vocoder. In the case of HNM vocoding scheme, the HNM analysis and excitation parameters of the CU are used by HNM vocoder.

9.5 Summary

We have presented a framework for generating synthetic listener vocalizations in unit selection speech synthesis from markup, using a combination of unit selection and signal modification techniques to generate synthetic vocalizations with more prosodic variety than what is contained in the recorded speech material. In this chapter, we first presented a markup specification to request vocalizations by the user. Secondly, we explained the drawbacks of simple unit selection algorithms and proposed an enhanced version of unit selection algorithm which is made use of signal modification techniques such as FD-PSOLA, MLSA vocoding, and HNM vocoding. Finally, we described our efforts to endow MARY TTS to synthesize listener vocalizations.

Chapter 10

Evaluation

The previous chapter proposed a realization algorithm to synthesize listener vocalizations according to given user requests. It was primarily aimed at *naturalness* and *appropriateness*. The signal modification techniques used in this algorithm aim to maintain a tradeoff between the quality of synthesized vocalizations and their appropriateness for the intended meaning requested by the user. This chapter evaluates the realization algorithm from these two perspectives.

This chapter is organized as follows. Section 10.1 describes our approach to evaluate the realization algorithm. Section 10.2 provides the details of corpus and annotation used in the evaluation experiments. Section 10.3 and Section 10.4 explain our two perception experiments conducted to evaluate the realization strategy. Section 10.5 summarizes the results.

10.1 Approach

This section describes our approach to evaluate the realization algorithm proposed in the previous chapter.

Figure 10.1 shows the symbolic notations used in the evaluation procedure. As shown in the figure, the proposed realization algorithm (in Chapter 9) first selects a best candidate unit (i.e. X) and an appropriate contour unit (i.e. Y) for a given user request. Next, it synthesizes a new vocalization (i.e. XY) by combining X's segmental form and Y's intonation contour using signal modification techniques; i.e. Y's intonation contour is transported onto the token X for the synthesis.

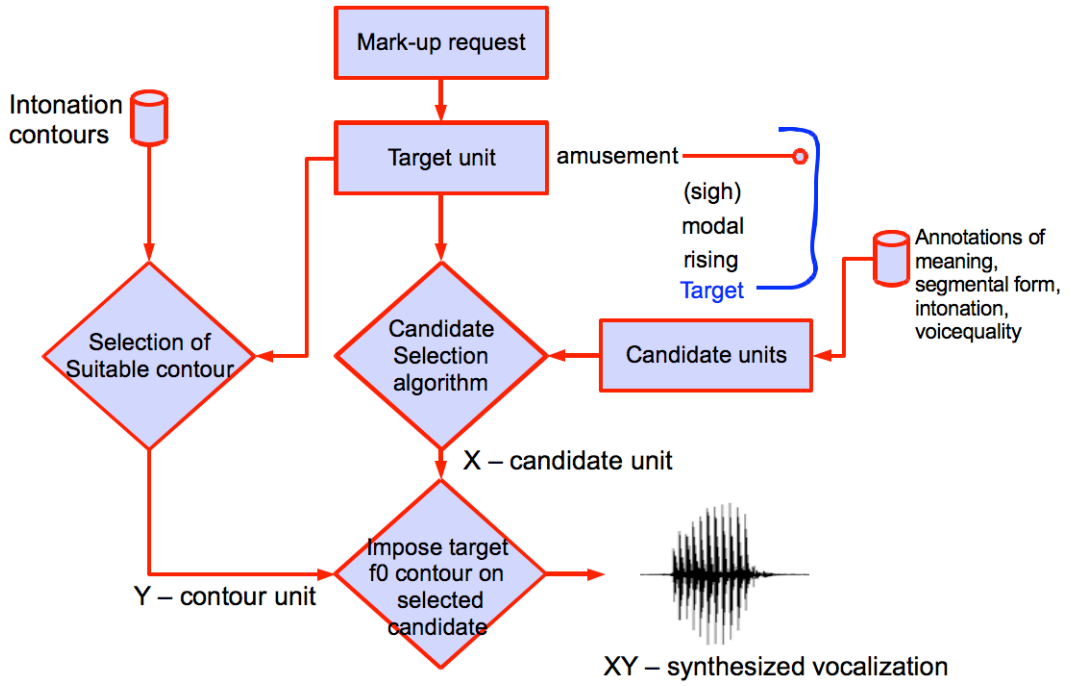


Figure 10.1: Symbolic notations used in the evaluation procedure: (i) *X* refers to the selected candidate unit; (ii) *Y* refers to the selected contour unit; (iii) *XY* refers to the synthesized vocalization that contains *X*'s segmental form and *Y*'s intonation contour.

Having described the notation, we now introduce the two obvious perspectives needed to evaluate the realization strategy: (i) the *naturalness* and the *influence on meaning* due to the usage of signal modification techniques for imposing *Y*'s intonation contour onto the token *X*; (ii) the *appropriateness* of synthesized vocalizations for the intended meaning – i.e. examining whether *XY* is more appropriate than *X* to signal the requested meaning.

We conduct two perception experiments to evaluate the above perspectives:

- **Perception experiment 1:** To evaluate the perceptual effects of applying different signal modification technologies when imposing intonation contours on vocalizations.

- **Perception experiment 2:** To evaluate whether the synthesized vocalizations convey a meaning closer to the intended meaning.

10.2 Database and annotation

In previous chapters, we have described our approach for data collection (in Chapter 6) and annotation of meaning and behavior (in Chapters 7 and 8). As described earlier, we collected data from four different voices - Poppy (cheerful), Prudence (pragmatic), Spike (aggressive), and Obadiah (gloomy). Among the recordings of four British English actors, this chapter uses Poppy's listener vocalizations (around 174 spontaneous vocalizations) for evaluation experiments.

In Chapter 8, we obtained annotation of meaning for a subset of 23 vocalizations as part of a study investigating the effect of segmental form and of intonation on the perceived meaning of listener vocalizations. About half of the vocalizations annotated differed in segmental form, but had approximately the same intonation contour (low and slightly falling); the other half had approximately the same segmental form (*yeah*) but varied in intonation contour. In a listening test, 20 subjects characterized each vocalization using the 11 meaning descriptors (*anger, sadness, amusement, happiness, contempt, solidarity, antagonism, certain, agreeing, interested*, and *anticipation*). In order to account for the expected inherent ambiguity of the listener vocalizations, descriptors were presented as scales, and subjects were asked to rate each vocalization on each of the scales.

The meaning of the unmodified vocalizations used in the present evaluation studies is based on the median of the subjective ratings for the intended meaning. The meaning and behavior annotation of these vocalizations can be found in Appendix B.

10.3 Experiment 1: effects of imposed F0 contours

The first perception experiment aims to evaluate the effects of applying different signal modification technologies when imposing intonation contours on vocalizations. The experiment is designed to address the following two research questions:

1. How good is the perceived naturalness of the resulting listener vocalizations after imposing an intonation contour (depending on the signal modification technology used)?

2. How does the meaning of the listener vocalizations change when cross-combining segmental form and intonation contour?

In the following subsections, we first describe the creation of stimuli for this perception experiment; secondly, we explain the listening test carried out for the evaluation; finally, we analyze the results of the study.

10.3.1 Creation of stimuli

To create stimulus material, three vocalizations were chosen through a semi-automatic process. We first applied a K-means algorithm to cluster the 174 vocalizations based on their intonation contours, using as criterion the second-order polynomial distance proposed by (Fujii, Kashioka, and Campbell 2003). Out of the resulting clusters, we selected a set of 17 vocalizations differing in segmental form. As a final step, we chose three vocalizations that were as different from one another as possible, in order to cover a reasonable range of segmental form as well as markedly different intonation contours.

As shown in Figure 10.2, the vocalizations have approximately the same length, and are voiced throughout. They are described as follows:

- mhm (A): low-falling contour with very narrow range;
- really (B): high-low jump with a very large F_0 range;
- oh (C): slow melodious fall from high to mid-range.

From these three original vocalizations, the synthesized stimuli were created as follows. Each of the three vocalizations was synthesized with each of the three intonation contours, using each of the three signal processing techniques (MLSA, FD-PSOLA and HNM). In total 27 synthetic stimuli are generated, out of these, nine are re-synthesized using the original intonation contour of the respective vocalization, and 18 are cross-synthesized using the other two intonation contours. We used these 27 synthetic plus the three original vocalizations as stimuli in the listening test. The original stimuli are included to provide reference data regarding meaning and naturalness ratings. The re-synthesized stimuli are included to provide insights in the effect of the signal modification algorithms as such, irrespective of a change in intonation contour. The cross-synthesized stimuli, finally, measure the effect of segmental form and intonation contour on ratings of meaning, and show the amount of degradation due to large modifications in intonation.

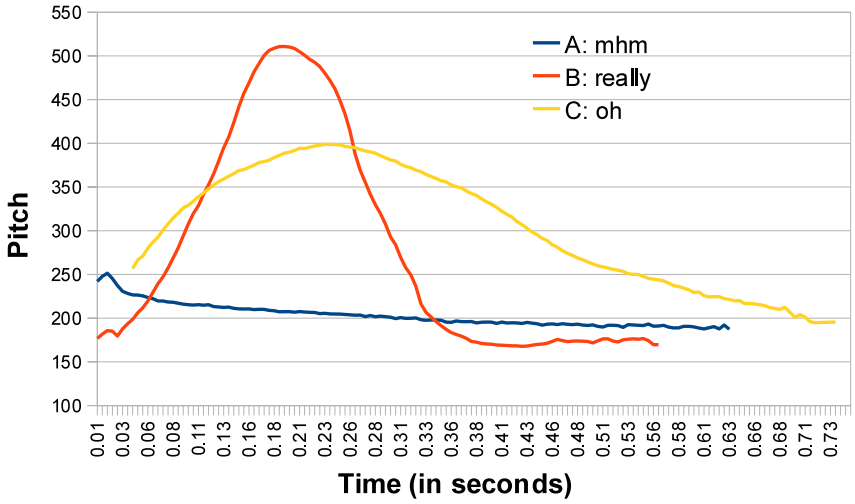


Figure 10.2: Intonations of the listener vocalizations

10.3.2 Listening test

A web-based listening test was conducted. Participants were presented with a task description, which included an explanation and examples of listener vocalizations, and made it explicit that synthetic vocalizations would be presented. Subjects were encouraged to use headphones and to adjust the playback volume before starting the test. The 30 stimuli were presented in an individually randomized order. Each stimulus, which could be re-played as often as the subject wished, had to be characterized using twelve five-point scales. The first scale measured the perceived naturalness (from 1 = completely artificial to 5 = completely natural). The remaining eleven scales were used to assess various aspects of meaning. This set of scales was described in Chapter 8. The meaning descriptors include seven unipolar scales: degree of *anger*, *sadness*, *amusement*, *happiness*, *contempt*, *solidarity* and *antagonism* (from "absolutely no X" to "pure uncontrolled X"), as well as four bipolar scales: *certain/uncertain*, *agreeing/disagreeing*, *interested/uninterested*, and *unexpected/anticipated*. Each of the eleven meaning descriptors was presented as a five-point scale. For each of the meaning scales (but not for the naturalness scale), subjects could tick a field "no real impression" if they felt it inappropriate to provide any scale value for a given meaning scale.

Nine subjects (five male, four female) participated in the test, most of whom were university staff from different language backgrounds. Given this heterogeneous pool of raters, any patterns with respect to meaning categorization are likely to be rather robust and not likely to be strongly culture-specific; however, it can only show a first trend. The annotation of meaning in the present test should be considered only as a first peek into the relative effects of segmental form and intonation in the perception of meaning. *Naturalness* and *meaning* ratings obtained for each of the stimuli in this perception study can be found in Appendix D.

10.3.3 Results and discussion

Naturalness

The naturalness ratings of the stimuli are shown in Figure 10.3. A clear pattern can be observed. First, the original stimuli are rated as most natural. Second, the stimuli which were re-synthesized with their own original intonation contour are slightly less natural. The third group of cross-synthesized stimuli, which are synthesized with a different vocalization's intonation contour, are substantially less natural. Within each group, HNM synthesis scores best, closely followed by FD-PSOLA, whereas MLSA scores clearly worse.

These findings confirm that the re-synthesis using FD-PSOLA and HNM introduce very few artifacts, whereas the quality already drops somewhat with re-synthesis using MLSA vocoding.

The fact that cross-synthesis is rated less natural than re-synthesis confirms the expectation that larger intonation modifications lead to more distortions. While this is established knowledge for the signal *modification* techniques FD-PSOLA and HNM, it might have been different in the case of MLSA vocoding. Given the fact that the signal is decomposed into a spectral envelope and an excitation and then vocoded from these representations, it would have been conceivable that this technology is more robust to larger F_0 changes. Our findings suggest that this is not the case.

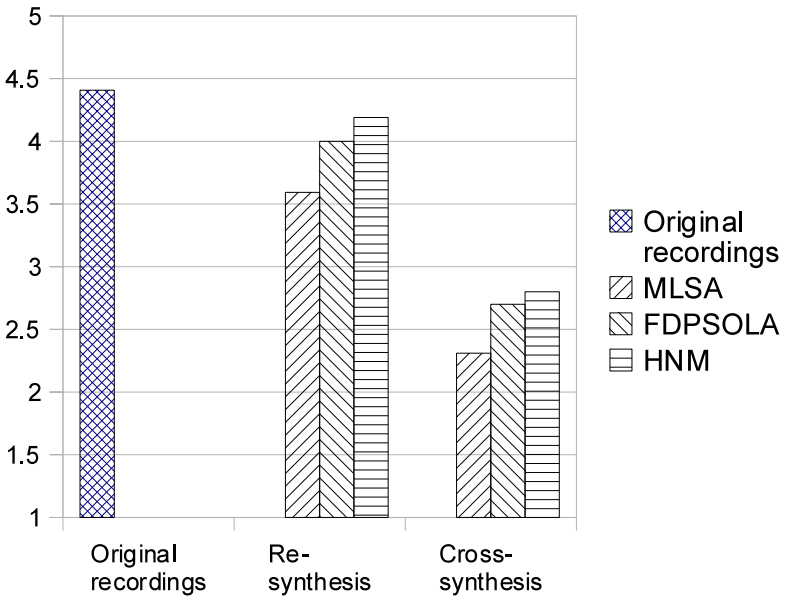


Figure 10.3: Naturalness ratings, with 1 = *completely artificial* and 5 = *completely natural*. Re-synthesis: the vocalization is synthesized with its own original intonation contour; Cross-synthesis: the vocalization is synthesized with the intonation contour taken from a different vocalization.

Meaning

In analyzing the ratings of meaning, we first looked at the "no real impression" ratings. Any scales for which more than half of the subjects indicated "no real impression" would be discarded; however, this criterion was never reached, so that all stimuli can be located on every scale.

The ratings of the meaning conveyed by the three original vocalizations *mhm_A*, *really_B* and *oh_C* can be seen in Figure 10.4 (a). First, it can be seen that all three vocalizations have received only moderate ratings on all scales, indicating that none of them was perceived as "pure uncontrolled" expression of any emotion. *mhm_A* was rated as somewhat sad, showing solidarity, uncertain and disagreeing. *really_B* was slightly amused and happy, showing solidarity, antagonistic, uncertain, clearly interested, and taken unawares. *oh_C*, finally, seems to have a rather diffuse meaning. According to the raters, it could express some sadness and contempt, but also amusement and happiness; it is high on solidarity but also shows some antagonism. In other

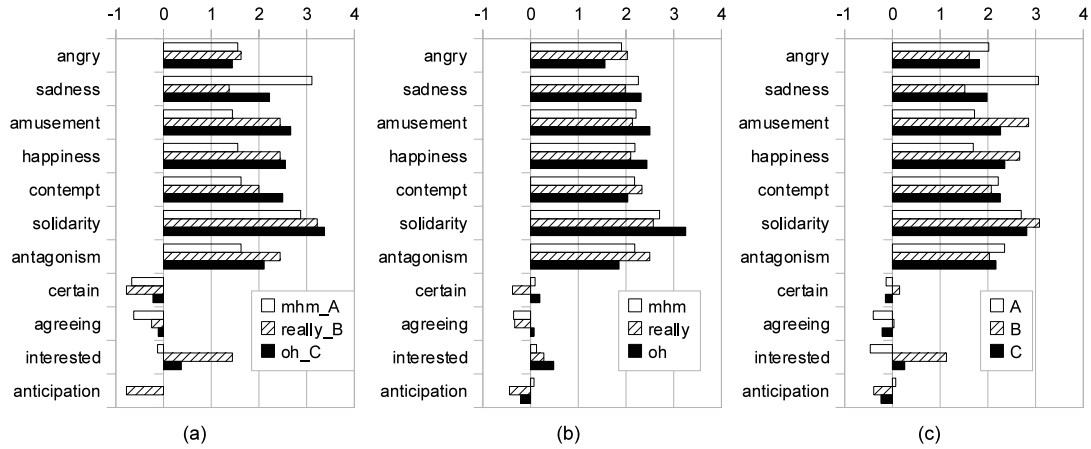


Figure 10.4: Average ratings of meaning, for (a) the three original vocalizations, (b) the three segmental forms (averaged over different intonation contours), and (c) the three intonation contours (averaged over different segmental forms). Scale values range from 1 (absolutely no X) to 5 (pure uncontrolled X) for the unipolar scales *angry*, *sadness*, *amusement*, *happiness*, *contempt*, *solidarity*, and *antagonism*, and from -2 to 2 for the bipolar scales, where -2 = *totally uncertain*, *totally disagreeing*, *totally disinterested* and *totally taken unawares*, and 2 = *totally certain*, *totally agreeing*, *totally interested*, and *anticipated events completely*.

words, *mhm_A* is a passive expression with negative valence but not directed against the interlocutor. *really_B* is a positive sign of interest and unexpectedness. *oh_C* seems to be an unspecific sign of solidarity with the interlocutor.

Figures 10.4 (b) and (c) show the extent to which these meanings are stable with the segmental form and with the intonation contour, respectively, when the other element is varied. In fact, it seems that the meaning differences due to segmental form (Figure 10.4 (b)) are rather small. *oh* is rated higher on solidarity than the other vocalizations, slightly lower on anger and antagonism, and higher on interest; *really* seems to express some antagonism, uncertainty, disagreement, and unexpectedness; *mhm* seems to have an element of disagreement but seems otherwise unspecific.

The rating patterns associated with the intonation contours, across vocalizations, are more conclusive (Figure 10.4 (c)). Contour *A*, the low and flat contour, is rated consistently high on sadness, low on amusement and happiness, and shows some disagreement and lack of interest. In contrast, contour *B*, the high-low jump, is low on

sadness but rather high on amusement, happiness and interest, and has an element of unexpectedness. The ratings for contour *C*, the high melodious fall to a mid range, show no clear pattern.

There were no systematic effects of signal modification method on meaning.

A detailed analysis of the interactions of segmental form and intonation shows interesting and partially unexpected interactions. For example, *really* is rated as somewhat angry with contours *A* and *C* but not with contour *B*; contours *B* and *C* are rated as more amused and happier with *mhm* and *oh* than with *really*; *really* is rated as quite contemptuous only when combined with contour *A*. Solidarity ratings for contours *A* and *C* are rather low with *really* but high with *oh*. *mhm* is rated as uncertain only with its original contour *A*; it is somewhat disagreeing with contours *A* and *C*, but is neutral or slightly agreeing with contour *B*. *really* is rated as highly interested with its original contour *B* but as quite uninterested with contour *A*.

These findings, even though the details may be questioned due to the small and heterogeneous set of listeners, seem to point out two important trends regarding the relative role of segmental form and intonation contour in determining the meaning of listener vocalizations. First, some but not all intonation contours seem to carry a specific meaning, which survives the combination of the contour with different segmental forms; similarly, some segmental forms seem to carry more specific meaning than others. Secondly, the meaning may change in unexpected ways when cross-combining segmental forms and meaning. For example, none of the ratings of the original vocalizations (Figure 10.4 (a)) allowed us to predict that *really* with the low and flat intonation contour *A* would convey anger and contempt.

10.4 Experiment 2: meaning-level appropriateness

The approach described in Chapter 9 allows us to synthesize arbitrary combinations of segmental forms and intonation contours. In this perception experiment we investigate whether the meaning of a synthesized vocalization can be modified towards an intended target meaning by imposing a suitable intonation contour onto an original listener vocalization.

Hypothesis: The realization approach makes the synthesized vocalizations convey a meaning closer to the intended meaning than an unmodified original in cases where no suitable match for the requested vocalization exists in the corpus.

In order to test the hypothesis, we use a pairwise comparison test where participants are requested to indicate which stimulus in the pair seems more appropriate for a given meaning. We chose this approach because we expect the pairwise presentation to make apparent more fine-grained distinctions than separate scale ratings. In particular, it is impossible to give an "undecided" answer.

10.4.1 Perception test

We prepared three types of stimuli – the original vocalizations of unit candidates (identified in the following as *X*), original vocalizations of contour candidates (henceforth, *Y*), and the synthesized vocalizations resulting from imposing *Y*'s contour onto *X* (henceforth, *XY*).

We selected 11 combinations of meanings and segmental forms to create stimuli. In choosing the combinations, we made sure that the segmental form with the intended meaning is not available in the corpus. Therefore, the unmodified vocalizations were expected not to convey the intended meaning. We tried to cover a reasonable range of segmental forms and meanings in the process of stimuli creation. This was not possible in all cases: for example, our data does not contain a single vocalization for *contempt*.

To evaluate the new approach, the three types (*X*, *Y*, *XY*) of stimuli were generated for the selected target combinations of segmental form and meaning. Table 10.1 shows the 33 stimuli with corresponding segmental forms, stylized intonation shapes, and meanings as previously annotated for *X* and *Y*.

intended meaning	***X***	***Y***	***XY***
amusement	*(sigh)* ○	yeah ●●	*(sigh)*
sadness	def'ly ●	yeah ●●	def'ly
anger	mh ○	yes ●	mh
happiness	yes ●	yeah ●●	yes
solidarity	mhmh ●	yeah ●●	mhmh
antagonism	def'ly ○	really ●●	def'ly
uncertain	yeah ●	*(sigh)* ●●	yeah
interested	gosh ●	yeah ●●	gosh
agreeing	mhmh ○	yeah ●●	mhmh
disagreeing	mhmh ○	yeah ●	mhmh
high anticipation[1]	gosh ○	def'ly ●	gosh

Table 10.1: Segmental form, intonation contour and previously annotated meaning of stimuli. *X*: original vocalization of unit candidate; *Y*: original vocalization of contour candidate; *XY*: synthesized vocalization, with segmental form from *X* and intonation contour from *Y*. def'ly: definitely. Meaning is represented using the following symbols.

○ : vocalization is not appropriate for the given meaning;

● : vocalization is somewhat appropriate;

●● : vocalization is very appropriate for the given meaning.

The pairwise comparison test is divided into three parts: *XY-X comparison*, *XY-Y comparison*, and *X-Y comparison*. The *XY-X comparison* is aimed to test the hypothesis, whereas the *X-Y comparison* is primarily a sanity check, verifying the expectation that *Y* is generally rated as better for the intended meaning than *X*. In addition, we carried out an *XY-Y comparison* to gain additional insight in the role of segmental form and intonation contours for the perceived meaning.

The three parts of the evaluation experiment were carried out through a web-based online perception test. Participants were presented with a task description, which included an explanation and examples of listener vocalizations, and made it explicit that synthetic vocalizations would be presented. Subjects were encouraged to use headphones and adjust the playback volume before starting the test.

Participants were asked which one among the two stimuli sounds more appropriate for a given meaning. For example, one question being asked in a comparison test was: *Which one of the following audio examples sounds more like "amusement"?* In total, 21 subjects participated in the online perception test.

10.4.2 Results and Discussion

The results of the listening test are shown in Figure 10.5.

Figure 10.5 (c) shows the *X-Y* comparison for the 10 usable stimulus pairs[1]. The ratings generally matched the expectation that *Y* should be perceived as closer to the respective meaning category than *X*. For *interest*, no clear preference between the two original vocalizations was found. This is not necessarily in conflict with the previous ratings (Table 10.1), where both *X* and *Y* were described as somewhat interested.

Figure 10.5 (a) shows the results that directly address our research question whether imposing a suitable intonation contour makes a vocalization more suitable for the intended meaning. Globally, the findings confirm the hypothesis. On average, the modified stimuli (*XY*) are preferred over the unmodified vocalizations (*X*) in 60% of the cases. This effect is statistically significant (Exact Binomial Test, two-sided $p_{binomial}(124, 207) < 0.01$). A closer look at Figure 10.5 (a) shows an inhomogeneous

[1]The comparison *X-Y* was included as a sanity check to make sure that subjects understood the terms used. For one meaning category, *high anticipation*, subjects nearly unanimously chose the *opposite* of the expected meaning. We conclude that they misunderstood the intended meaning (viz., as something that was highly anticipated and predictable), and therefore removed the data from further analysis.

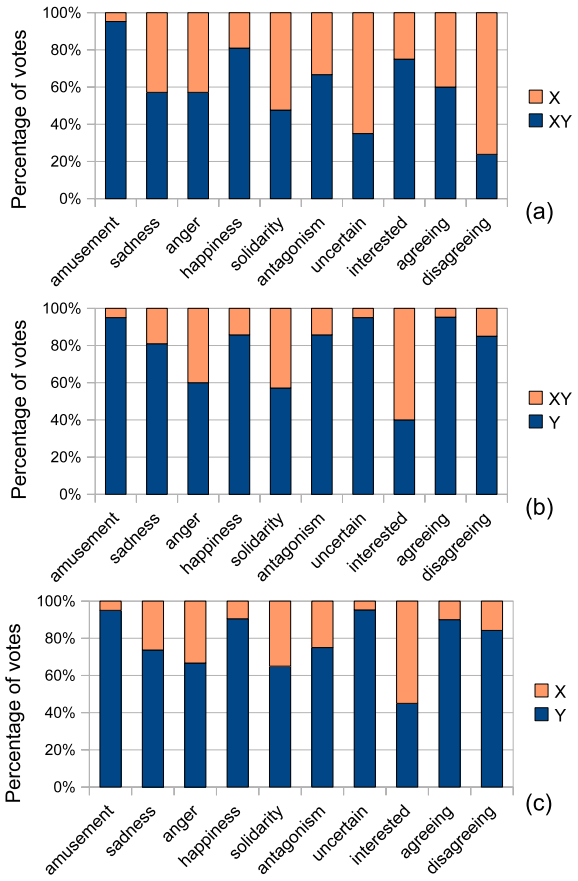


Figure 10.5: Percent of vocalizations rated as more appropriate for the given meaning, comparing (a) original *X* with synthesis *XY*; (b) original *Y* with synthesis *XY*; and (c) originals *X* and *Y*.

picture, however. For *amusement*, *happiness* and *interested*, the intonation contour improved the recognition as the intended meaning category significantly (Exact Binomial Test, $p < .05$ or better). A significant effect in the opposite direction is found for *disagreeing*: here, the combined stimulus is rated as consistently *less* appropriate than the unmodified original. Most of the remaining pairs showed a trend towards a preference for the *XY* vocalization, but the effects did not reach significance individually.

Regarding the comparison between the modified vocalization *XY* and the vocalization *Y* from which the respective intonation contour was taken (Figure 10.5 (b)), we find a very strong global effect of *Y* being preferred over *XY* (Exact Binomial Test,

two-sided $p_{binomial}(45, 205) < 0.001$). This effect is also significant for all individual pairs except for *anger*, *solidarity* and *interested* (Exact Binomial Test, $p < .01$ or better).

Figures 10.5 (b) and (c) show a nearly identical pattern: obviously, imposing the approximate intonation contour was not sufficient to reach the appropriateness of the original.

These findings appear to confirm, with qualifications, the hypothesis tested in this experiment: that a vocalization's suitability for a certain meaning can be improved by imposing on it an intonation contour taken from a "good example" for the intended meaning. The finding is particularly strong with the stimuli using the most extreme intonation contour in the set: out of the four stimuli using the high-fall contour as a target (see Table 10.1), three are rated as significantly more appropriate: *amusement*, *happiness*, and interested (see Figure 10.5 (a)). It may be that for the remaining stimuli, the intonation contours of source *X* and contour target *Y* were actually too similar, so that the perceptual effect on *XY* may have been too subtle.

10.5 Summary

This chapter has presented two evaluation experiments. The first listening experiment investigated the perceptual effects of imposing one intonation contour onto another vocalization using different signal modification techniques. The second one was aimed to evaluate whether the synthesized vocalizations convey a meaning closer to the intended meaning.

In the first perception experiment, we have experimentally investigated the perceptual effects of imposing intonation contours onto a small selection of different vocalizations, using three state-of-the-art signal modification techniques: MLSA vocoding, FD-PSOLA and HNM. Our findings indicate that the drop in naturalness seems strongest for MLSA and smallest for HNM and FD-PSOLA; naturalness degrades substantially when imposing intonation contours that are very different from the original contour, but at least for HNM and FD-PSOLA stays high when re-synthesizing the original contour. In line with the literature, we expect this to be a continuous effect, in the sense that smaller changes to the intonation contour should also lead to smaller degradations.

Regarding the meaning of listener vocalizations, we have found distinguishable meanings of some, but not all, segmental forms and intonation contours. Unexpected interactions were observed, where certain configurations of segmental form and intonation caused a perceptual impression that was not predictable from the individual meanings of segmental form and intonation separately. This means that, when synthesizing from meaning-level markup, caution seems to be of order when combining segmental forms and intonation contours.

The second perception experiment evaluated the realization algorithm, which was presented in Chapter 9, in the unit selection domain for increasing the range of vocalizations that can be synthesized with a given set of recordings. The algorithm takes a vocalization with the intended segmental form and imposes an intonation contour from another vocalization with the intended meaning onto it using FD-PSOLA. In the second listening test, the modified versions were rated as significantly more appropriate for the intended meaning category than the unmodified vocalizations. This appears to confirm that the algorithm can make available for use in synthesis combinations of segmental form and meaning that have not been recorded.

The effect was clearest in the cases where an extreme intonation contour was imposed. It may be suboptimal to favour target contours that are as similar as possible to the source contour; it remains to be seen to what extent the benefits of using a markedly different contour outweigh the cost of more perceivable distortions. Alternatively, if the annotation of vocalizations included the extent to which segmental form, voice quality and intonation conveyed a certain meaning, we could use only contours that are informative. For the moment, however, we do not see how to obtain such annotation with reasonable effort.

Part III

Reflection

Chapter 11

Applications and usage

In previous chapters we have presented our investigation on synthesis of listener vocalizations. The work was mainly focussed on data collection, annotation and realization strategies to generate appropriate and natural listener vocalizations. This chapter discusses some interesting collaborative and ongoing work that is enabled by the research involved in this thesis.

In Section 11.1 we describe the integration of synthesis of listener vocalizations into the SEMAINE framework. Section 11.2 provides a brief overview of recent investigations on multimodal listening behavior of an ECA. Finally, Section 11.3 briefly mentions the ongoing research work on 'listening robots' as part of another research project, which is enabled by this thesis work.

11.1 Listening SAL characters

An obvious application area for the synthesis of listener vocalizations is the SEMAINE project's Sensitive Artificial Listeners (SALs). As we described in Chapter 1, integrating the ability to generate listener vocalizations into the SEMAINE framework is one of the objectives of this thesis. We described the SEMAINE framework in Chapter 4. This section briefly discusses technicalities involved in the realization procedure of listener behavior in the SAL framework. In addition, we also describe our approach to integrate the ability to generate appropriate listener vocalizations in the multimodal generation framework.

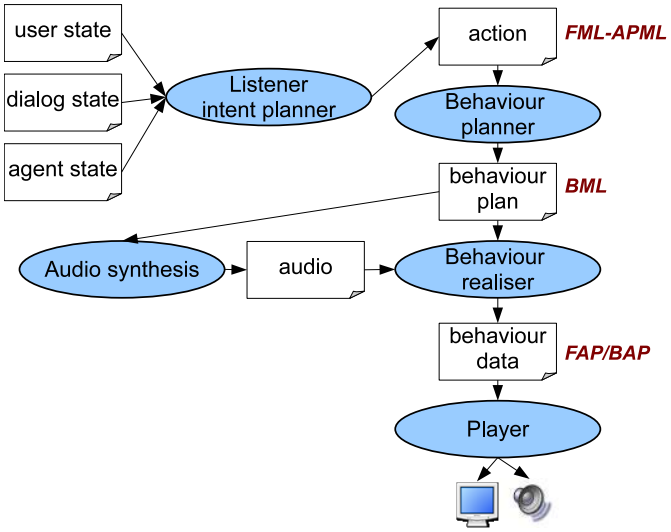


Figure 11.1: The realization procedure of listener behavior in SAL framework

As shown in Figure 11.1, a *listener intent planner* generates the agent's listener behavior based on the user's behaviors; for example, a head nod or a variation in the pitch of the user's voice will trigger a listener response with a certain probability. The *behaviour planner* takes as input both the agent's communicative intentions specified by the FML-APML (Functional Markup Language - Affective Presentation Markup Language) language (Heylen et al. 2008) and some of the agent's characteristics. The main task of this component is to select, for each communicative intention to transmit, the adequate set of behaviors to display. All possible sets of behaviors for a given communicative intention are defined in a *lexicon*. For example, the listener response (i.e. backchannel) that wants to transmit the communicative function "agreement" is generated by the Listener Intent Planner in the following FML[1]:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
 <fml-apml>
  <fml xmlns="http://www.mindmakers.org/fml" id="fml1">
    <backchannel id="b1" type="agreement"
              start="0.0" end="1.5" importance="1.0"/>
  </fml>
 </fml-apml>
```

[1] http://semaine.opendfki.de/wiki/FML

From the FML tag, the behavior planner automatically generates a multimodal set of behavioral signals written in Behavior Markup Language (BML)[1], taking into account the agent's personality and its lexicon. For example, for Poppy, the happy and optimistic character, the behavior planner could generate the following BML:

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<bml xmlns="http://www.mindmakers.org/projects/BML">

<head id="backchannel-0" start="0.00" end="1.50"
                   direction="RIGHT" type="NOD">
  <description level="1" type="gretabml">
   <reference>head=head_nod</reference>
      <FLD.value>0.00</FLD.value>
      <OAC.value>0.80</OAC.value>
      <PWR.value>0.20</PWR.value>
      <REP.value>0.00</REP.value>
      <SPC.value>0.90</SPC.value>
      <TMP.value>0.80</TMP.value>
      <preference.value>0.90</preference.value>
  </description>
</head>

<face id="backchannel-1" start="0.00" end="1.50"
                  shape="flat" type="MOUTH">
  <description level="1" type="gretabml">
   <reference>eyes=smile</reference>
      <FLD.value>0.50</FLD.value>
      <OAC.value>0.80</OAC.value>
      <PWR.value>0.20</PWR.value>
      <REP.value>0.00</REP.value>
      <SPC.value>0.90</SPC.value>
      <TMP.value>0.80</TMP.value>
      <preference.value>0.70</preference.value>
  </description>
</face>
```

[1] http://semaine.opendfki.de/wiki/BML

```
<speech id="backchannel-2" language="en-GB" text="yes"
                  type="application/wav" voice="dfki-poppy">
  <vocalization xmlns="http://mary.dfki.de/2002/MaryXML/"
     intonation="rising"
     meaning="agreeing"
     name="yeah"
     voicequality="breathy"/>
   <description level="1" type="gretabml"/>
      yeah
 </speech>
</bml>
```

A suitable behavior of different modalities such as *face, head*, and *speech* is generated according to the agent's lexicon. According to the above BML example, Poppy smiles and nods while saying "yeah" in a friendly way to show "agreement". This behavior plan contains the markup information to synthesize listener vocalizations (highlighted text in the example). The audio synthesis module takes the corresponding speech modality markup and realizes appropriate listener vocalization using MARY TTS. The audio of synthesized listener vocalization and corresponding durations and timing information for its phone segments (i.e. phonetic labels) are sent to the behavior realizer in order to realize them jointly with the visual behavior of SAL agent with proper lip synchronization.

In order to plan the agent's behavior for a given FML, the behavior planner maintains a lexicon file which contains a set of behaviors with alternative signals for different modalities. For example, the following are alternative vocalizations for the meaning 'agreement'; the behavior planner picks up one among the alternatives based on previously generated signals.

```
<signal id="4" name="text" modality="speech" content="yes"
   intonation="rising" voicequality="tense" meaning="agreeing">
   <alternative name="text" content="yes" intonation="rising"
       voicequality="modal" meaning="agreeing"
             probability="0.2"/>
   <alternative name="text" content="tsyeah" intonation="rising"
       voicequality="modal" meaning="agreeing"
             probability="0.1"/>
   <alternative name="text" content="tsright" intonation="rising"
       voicequality="modal" meaning="agreeing"
             probability="0.1"/>
   <alternative name="text" content="right" intonation="rising"
       voicequality="modal" meaning="agreeing"
             probability="0.1"/>
   <alternative name="text" content="alright" intonation="rising"
       voicequality="modal" meaning="agreeing"
             probability="0.1"/>
   <alternative name="text" content="yeah" intonation="rising"
       voicequality="breathy" meaning="agreeing"
             probability="0.1"/>
   <alternative name="text" content="yeah_that's_true"
       intonation="rising" voicequality="modal"
             meaning="agreeing" probability="0.1"/>
</signal>
```

The behavior realizer module generates the animation of the SAL agent in MPEG-4 format. The input of the module, which is specified by the BML language, contains the text to be spoken and/or a set of nonverbal signals to be displayed. In addition, the audio synthesis module provides the synthetic speech with the list of phonemes and their respective durations. This information is used to compute the lips movements. Each BML tag is instantiated as a set of key-frames that are then smoothly interpolated. Facial expressions, gaze, gestures, torso movements are described symbolically in repository files. Finally, the player receives the animation, which is defined by Facial Animation Parameters (FAPs) and Body Animation Parameters (BAPs), and plays it in a graphic window.

11.2 Evaluation of multimodal listener responses

This thesis enabled a joint evaluation study in which we analyzed the perception of multimodal backchannels of ECAs (Bevacqua et al. 2010). The aim of the evaluation consisted in improving and extending the backchannel lexicon by introducing multimodal signals in a consistent manner, that is by taking into account the interaction between visual and acoustic cues and the meaning that they can convey when displayed together. To this purpose, we asked subjects to judge a set of multimodal signals performed by the 3D agent GRETA (Niewiadomski et al. 2009). We considered in this perceptual evaluation the twelve meanings: *agreement, disagreement, acceptance, refusal, interest, not interest, belief, disbelief, understanding, not understanding, liking, disliking*. The signals were context-free, that is without knowing the discursive context of the speaker's speech. To create videos we selected 7 visual signals and 8 audio signals (7 vocalizations plus silence). The visual signals were chosen from previous evaluation studies (Heylen et al. 2007; Bevacqua et al. 2007): *raise eyebrows, nod, smile frown, raise left eyebrow, shake*, and *tilt&frown*. The vocalizations were selected using an informal listening test. Initially, three participants assigned each of the 174 vocalizations produced by the speaker to one of the 12 meanings used in this experiment. We then selected the seven stimuli which seemed least ambiguous for their respective meaning, in order to cover a reasonable range of different vocalizations. The tokens of selected vocalizations are *ok, ooh, gosh, really, yeah, no, m-mh* and *silence*. We generated 56 multimodal signals as the combinations of the visual and acoustic cues selected.

As part of the study, we tested the following three hypotheses:

- **Hp1:** the strongest attribution of a meaning will be conveyed by the multimodal signals obtained by the combination of visual and acoustic cues representative of the given meaning.
- **Hp2:** in some occasion, multimodal signals convey a meaning different from the ones associated to the particular visual and acoustic cues when presented on their own.
- **Hp3:** visual and acoustic signals that have strongly opposite meanings are rated as nonsense: like *nod+no, shake+ok, shake+yeah*.

55 participants (22 women, 33 men) with a mean age of 31.5 years, mainly from France (33%), Italy (18%), accessed anonymously to the evaluation through a web browser. The first page provided instructions, the second collected demographic information. Then the multimodal signals were played one at a time. Participants used a bipolar 7-points Likert scale: from -3 (extremely negative attribution) to +3 (extremely positive attribution). The evaluation was in English.

The first hypothesis has been only partially satisfied. Results showed that the strongest attribution for a meaning is not always conveyed by the multimodal signals obtained by the combination of visual and acoustic cues representative of the given meaning. That means that the meaning conveyed by a multimodal backchannel cannot be simply inferred by the meaning of each visual and acoustic cues that compose it. It must be considered and studied as a whole to determine the meaning it transmits when displayed by virtual agents. Moreover, we found that some multimodal signals convey a meaning different from the ones associated to the particular visual and acoustic cues when presented on their own (Hp2). The evaluation showed also that multimodal signals composed by visual and acoustic cues that have strongly opposite meanings are rated as nonsense. As expected *nod+no*, *shake+yeah*, *shake+ok* and *shake+really* were rated as senseless. What is more, a high attribution of nonsense does not necessarily exclude the attribution of other meanings. Thus, the high nonsense signal of *shake+yeah* was also highly judged as showing disbelief. A possible explanation would be that these signals might be particularly context depend. This evaluation gave us a better insight about several multimodal backchannels and the meaning they convey. The results have been used to enrich and expand the backchannel lexicon of our virtual agent.

11.3 Listening robots

The integration of listener behavior into NAO robots is another interesting application which is being attempted as part of the ALIZ-E project[1] – which is aimed to develop embodied cognitive robots for believable any-depth affective interactions with young users over an extended and possibly discontinuous period. This ongoing project

[1] http://aliz-e.org

uses event-based component integration (Kruijff-Korbayová et al. 2011), where Urbi SDK (Baillie 2005) is used as the middleware. The integrated system consists of several components, such as *dialog manager, speech recognizer, natural language understanding, natural language generation*, and *text-to-speech*. Each of the available components communicates with each other in the system using the Urbi middleware.

In order to improve interactiveness and spontaneous nature of the robotic interaction with young children, the listener's behavior is being integrated into the ALIZ-E project demonstration system. The initial system is a preliminary system which uses a simple mapping between *speech acts* that are generated by the dialog manager and the vocalizations to be synthesized. For example, the speech act *providePositiveFeedback* can be realized by either a happiness *wow* vocalization or an interested *gosh* vocalization. This type of mapping is used to find *which* vocalization has to be realized at runtime.

The recent integration of MARY TTS in the ALIZ-E framework has enabled generating synthesized vocalizations through the NAO robot's speakers. However, appropriate robotic gestures for the corresponding vocalizations have to be investigated in future.

11.4 Summary

This chapter has described some of the applications and ongoing research work in human-machine interaction that are enabled by this thesis work. We explained how the listening behavior is incorporated into Sensitive Artificial Listeners (SAL) characters in order to turn them into 'listening SAL' agents. A collaborative research work on multimodal listening behavior of ECAs was briefly discussed; we present results of a perception study on multimodal (i.e. visible and audible acts) behavior of an ECA. Finally, we very briefly describe the current efforts to achieve 'listening robots' by incorporating the thesis work into the ALIZ-E project.

Chapter 12

Conclusions and future work

This chapter draws conclusions of the work in this thesis. It summarizes our work showing what we have achieved. We also present how this thesis can be further extended in order to achieve interactive speech systems.

12.1 Achievements

This thesis represents the first attempt to incorporate the ability to synthesize natural listener vocalizations in a full-scale speech synthesis system. The main achievements of the investigation are as follows: (i) collection of natural listener vocalizations; (ii) annotation of meaning and behavior; (iii) realization of appropriate listener vocalizations based on markup requests. This thesis has been written in the context of the SEMAINE project; therefore, an additional objective of the thesis has been to (iv) integrate our work into the SEMAINE framework in order to add listening capabilities to the SAL characters.

(i) Collection of natural listener vocalizations

We described a method to collect natural listener vocalizations. According to this method, we recorded dialogue speech between actors and their dialogue partners. We instructed the actors to take predominantly a listener role, however, the actors were allowed to take the speaker role in order to maintain a natural dialogue. Our student assistants extracted listener vocalizations from the natural dialogue, and provided

'single-word' description for each of the listener vocalizations. In this way, we collected 1080 listener vocalizations from around six hours of dialogue speech with a German professional actor, while a two-hour dialogue speech corpus from four British English actors contains around 480 listener vocalizations. In this method, we did not have much control over the range of response tokens used in the dialogue and their acoustic variability. However, this method of data collection seemed to be successful to collect natural listener vocalizations.

(ii) Annotation of meaning and behavior

An exploratory annotation study was conducted in order to find more suitable meaning and behavior descriptors to represent listener vocalizations. We used Baron-Cohen et al. (2001)'s epistemic states and Scherer (2005)'s Geneva Emotion Wheel (GEW) affective categories as starting sources to describe meaning of listener vocalizations. In this study, we finally used 37 meaning descriptors to annotate German listener vocalizations; among them, 11 meaning descriptors were frequently used; i.e. each of those were represented at least 5% of the vocalizations used in the study. However, the inter-rater agreement on a small sub-corpus was low, which signifies that the list of meaning descriptors needs to be consolidated. Different methods were used to annotate behavior properties such as segmental form, intonation and voice quality. We used a 'single-word' description to represent segmental form. We proposed a semi-automatic procedure, which includes a hierarchical agglomerative clustering of F0 contours, to annotate intonation contours with symbols based on their shapes. For voice quality annotation, we used Laver (1991)'s descriptors, which showed higher inter-rater agreements. The exploratory annotation was found to be a tedious and time-consuming procedure, however essential for finding a suitable set of meaning and behavior descriptors.

A systematic study of meaning annotation was conducted to annotate meaning *appropriateness* on scales. We first consolidated the list of meaning descriptors to 11 scales: *anger, sadness, amusement, happiness, contempt, solidarity, antagonism, (un)certain, (dis)agreeing, (un)interested*, and *(high/low)anticipation*. We then conducted a perception test on a set of selected stimuli that includes representative vocalizations from the British English corpus. This multidimensional meaning annotation

was carried out as a web-based perception test in order to facilitate participation of several subjects from different parts of the world. This multidimensional annotation approach helped us to find *typical* impressions of several raters on meaning appropriateness of vocalizations. We observed inherent ambiguity in listener vocalizations, where each of the listener vocalizations tends to convey multiple meanings at the same time. In addition, this experimentation permitted us to investigate the relevance of segmental form and intonation contours on the meaning of listener vocalizations. The evidence indicated that the intonation contour is highly relevant for signaling meaning when compared to the phonetic segmental form.

(iii) Realization: unit selection and signal modification

A synthesizer should be capable of generating appropriate and high quality listener vocalizations. It should also have the ability to synthesize a broad range of vocalizations to communicate different intentions with different kinds of acoustic properties. In a simple unit selection algorithm, where a finite set of recorded listener vocalizations is available, synthesis quality is high, but the acoustic variability is limited. As a result, many combinations of segmental form and intended meaning cannot be synthesized. In other words, it can not support synthesis of 'unseen' vocalizations.

To overcome this limitation, we developed an enhanced version of the unit selection algorithm with imposed intonation contours. This algorithm first selects the suitable candidate units and intonation contours separately for a given target based on cost based selection scheme; secondly, the best candidate and contour pair is selected using another cost function; finally, in order to improve the *appropriateness* of synthesized vocalization for the given meaning, the target contour is transported onto the candidate unit using signal modification algorithms, such as FD-PSOLA, HNM vocoding, and MLSA vocoding. The known limitation of the new approach, which is based on unit selection and signal modification techniques, is that we can not synthesize segmental forms that are not available in our corpus. For example, if the user requests *oh-yeah* vocalization and if that token is not available in the recorded corpus, we can not synthesize it. Nevertheless, the approach can improve the *appropriateness* of the synthesized vocalization towards the intended meaning.

The MARY TTS framework was extended not only to implement the above algorithm, but also to generate listener vocalizations based on an XML request. The TTS system stores the recorded audio of each vocalization together with phone segment labels and features representing the segmental form, intonation, voice quality and possible meanings of the vocalization, as annotated previously. At run-time synthesis, MARY TTS uses the proposed approach to synthesize appropriate listener vocalizations for a given markup.

Evaluation

The proposed strategy was aimed to maintain a tradeoff between appropriateness and naturalness to synthesize both appropriate and high quality vocalizations at the same time. We evaluated, on the one hand, the perceptual effects of signal modification techniques on the *naturalness* and the *meaning* of synthesized vocalizations; on the other hand, we evaluated whether the unit selection process has an impact on the *appropriateness* of synthesized vocalizations. In the first perception experiment, prosody modification technique based on HNM vocoding showed relatively good performance, whereas MLSA vocoding performance showed to be low. The cross combination of segmental form and intonation contour has shown no clear patterns on the perceived meaning. In the second listening test, the modified versions were rated as significantly more appropriate for the intended meaning category than the unmodified vocalizations. This appears to confirm that the algorithm can make use of combinations of segmental form and meaning that have not been recorded. The effect was clearest in the cases where an extreme intonation contour was imposed. It may be suboptimal to favor target contours that are as similar as possible to the source contour; it remains to be seen to what extent the benefits of using a markedly different contour outweigh the cost of more perceivable distortions.

(iv) Integration into the SEMAINE framework

We have briefly described our efforts to integrate the ability to synthesize listener vocalizations into the SEMAINE framework. A *lexicon* was maintained to describe suitable alternative behaviors for an intended meaning. The probabilities of the alterna-

tives are continuously updated based on recently generated signals. A signal that has highest probability at a certain time can be triggered in order to produce a listener response, which can be synthesized with the SEMAINE framework using MARY TTS. The synthesized audio and corresponding phoneme durations are used to compute lip movements before rendering the audio with visual behavior of the SAL agent.

We conducted a perception study in order to understand how listener vocalizations influence and/or modify the meaning of visual listener responses. We investigated perceptual effects of ECA's multimodal listening behavior when we cross combine: (i) different visual expressions such as *head-nod, head-shake, smile* and *frown*; and (ii) listener vocalizations such as *uh-huh, yeah, gosh* and *really*. The results indicate that the strongest attribution for a meaning is not always conveyed by the multimodal signals obtained by the combination of visual and acoustic cues representative of the given meaning. Moreover, we found that some multimodal signals convey a meaning different from the ones associated to the particular visual and acoustic cues when presented on their own. That means that the meaning conveyed by a multimodal listener response cannot be simply inferred by the meaning of each visual and acoustic cues that compose it. It must be considered and studied as a whole to determine the meaning it transmits when displayed by virtual agents.

12.2 Future work

This thesis has investigated only one among several research issues in order to achieve *interactiveness* in conversational agents. This section describes possible future directions.

Voice quality conversion techniques

The meaning of a listener vocalization is influenced by behavior properties such as *intonation* and *voice quality*. The realization procedure presented in this thesis is focussed on imposed intonation contours only. Nonetheless, the evaluation shows reasonably good improvements in *appropriateness* on the perceived meaning of synthesized vocalizations. However, the inclusion of voice quality conversion techniques,

such as *modal to breathy*, *modal to creacky*, *breathy to lax* etc. conversions, is expected to improve the results of the procedure.

Synthesis of 'unseen' segmental forms

The inability to synthesize tokens that are not available in the recorded corpus is one of the limitations of the approach proposed in this thesis. To overcome such limitation, we have to investigate possibilities to synthesize new vocalizations based on phonetic descriptions of available listener vocalizations. For example, concatenating phones of available vocalizations to generate new vocalizations is a possibility. However, it is not an easy task due to extreme prosodic variations of listener vocalizations.

Investigating parametric speech synthesis techniques to generate new tokens would be another possible direction. Hidden Markov Model (HMM) training methods to model phonetic descriptions of vocalizations have to be investigated. However, such training methods may suffer with data scarcity problem in this case.

Taking context into account

Contextual appropriateness is not investigated as part of this thesis. As described in Chapter 2, the meaning and the behavior of a listener vocalization certainly depend on the speech context produced by the interlocutor. However, the contextual behavior such as audible and visible acts of the interlocutor is not considered in this research. Finding contextually appropriate listener vocalizations for a situation would be another possible direction for the future work. Such investigation could include several stages such as: (i) recording context-sensitive conversational databases which will be useful for synthesis; (ii) annotation of contextual meaning appropriateness; (iii) identification of best features to represent context; (iv) investigating machine learning techniques to find suitable meaning for the observed context.

Vocalizations when speaking

Until now, we discussed the usage of the vocalizations when the agent is listening. In spontaneous conversations, we use vocalizations such as interjections while speaking,

as well. For example, sentences like *Oh! My dear daddy* and *Wow! It is wonderful*. An emotionally colored conversational speech synthesis system is expected to synthesize such vocalizations in a way that is consistent with their context. This topic raises interesting research questions such as 'how to realize the behavior of a vocalization such that it is compatible with its contextual speech?'. In other words, 'how to maintain contextual appropriateness with his/her own speech?'. Investigation of such speaking behavior is another challenging topic for the future.

Listening behavior of ECAs

One of the most desirable characteristics of an Embodied Conversational Agent (ECA) is the capability of interacting with users in a human-like manner. Such an ECA should be able to socially interact with human users over extended periods of time. Multimodal listening behavior is expected to increase the *believability* in interaction, which could lead to long-term interaction. While listening to a user, ECA should be able to provide listener responses through all those modalities that human beings use to communicate (speech, facial expression, gesture, head movement and so on). Although we evaluated perceptual effects when different visual expressions and listener vocalizations are realized together, the obtained results indicate that further research is needed for thorough understanding of ECA's multimodal listening behavior.

Gestures of listening robots

In human-robot interaction, gestures of listening robots can be considered as unexplored. In order to integrate listening behavior into robots, suitability of several robotic gestures such as hand and head gestures for different meanings have to be investigated. In addition, we have to study the perceptual effects of the gestures with listener vocalizations when they are realized together.

Appendix A

Web-based perception experiment

The process of meaning annotation is a crucial step in this thesis. This appendix provides a short description on the web-based perception study that is used to get meaning ratings from several subjects.

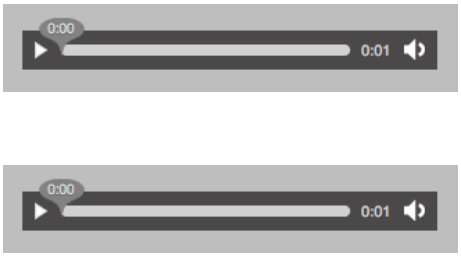
Perception of listener vocalisations (Page 1)

This research is part of a larger project, an EU funded project called SEMAINE, that has the goal of developing a fully automatic computer system that will be able to hold a sustained and convincing conversation with the user. The challenge for computer developers is that so much of human conversation is actually non-verbal. The subtle changes in tone of voice or gesture and shifts in posture or eye contact all help to regulate our conversations with other humans – but of course this all comes so naturally to us that we are seldom aware of it.

The purpose of this study is to understand the meaning behind different listener vocalizations like *mhm, right, yeah, uh-huh* etc. In order to do this we are gathering opinions of people on different listener vocalizations.

Example of listener vocalizations with dialog context :

The same listener vocalizations without dialog context :



Please check the following before you proceed:

- If you cannot see audio players above, you could try a QuickTime plugin
- We recommend Firefox ($>$ 3.5).
- Make sure that you have adjusted to proper volume of your audio device (headphones)

As this perception study is aimed to get user ratings on meaning of listener vocalizations, you will be asked to play audio file in order to listen to each vocalization and you have to rate meanings on the basis of 5-point scale. You can listen to the audio as many times as you need. The set of meaning categories include 'anger', 'amusement', 'solidarity' (positive attitude to interactant), 'antagonism' (negative attitude to interactant) etc.

Your participation is entirely voluntary, you do not have to participate, you are free to withdraw at any time without having to provide any further explanation and there will be no negative implications as a result of your withdrawal. If you decide to withdraw from the experiment, any recording already made will be erased (to ensure your data is erased send an e-mail to us).

Please complete the whole evaluation and it will take about 1 hour to complete. This evaluation can be done in different sessions. You can stop this session at any time. Please keep a note of the current webpage link (redirected one) and the email address you have used to register in case you do not complete the evaluation in this session. If you enter the same Email address, it will take you to the exact position where you stopped in your previous session. Your participation is warmly appreciated. Thank you for your time.

Please enter your E-Mail address to continue with your participation :

(This is a research study. Your email address will not be used except to keep track of where you have reached. Feel free to ask any questions you may have about the study in an email to the experimenters.)
Ethical approval for this research was granted by the Queen's University Belfast Psychology Research Ethics Committee.

Participant consent form (Page2)

1. I agree to participate in this research. □
2. This agreement is of my own free will. □
3. I am aware of contact e-mails to ask any questions about the study. □
4. I realise that I may withdraw from the study at any time, without giving a reason. □
5. I realise that I may withdraw any recorded data from the study at any time without giving a reason. □
6. I have been given full information regarding the aims of the research, and have been given information with the researcher's name and a contact number and address if I require further information. □

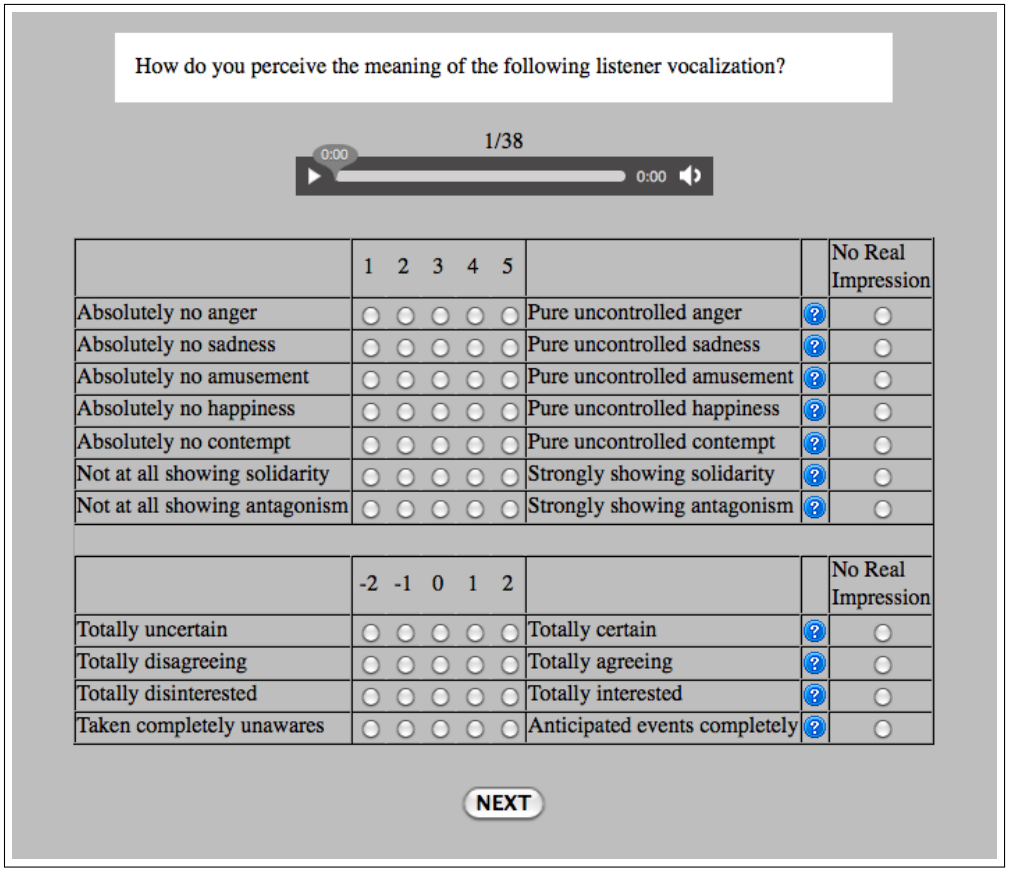
7. All personal information provided by myself will remain Confidential as outlined in the Participant Information Sheet. □

8. I agree that the researchers may store the data in question indefinitely, may share them with the other SEMAINE researchers and may use them for research purposes as they deem appropriate. □

9. I agree to allow other research teams (outside SEMAINE) to have access to the data if the researchers consider it appropriate. (Allowing access to other research teams will not include giving them permission to distribute the recordings. The researchers may ask other teams to pay costs associated with providing access, but may not sell the material or charge commercial sums for it.) □

□ **I agree to participate in this research.**

Perception Test (from Page 3)

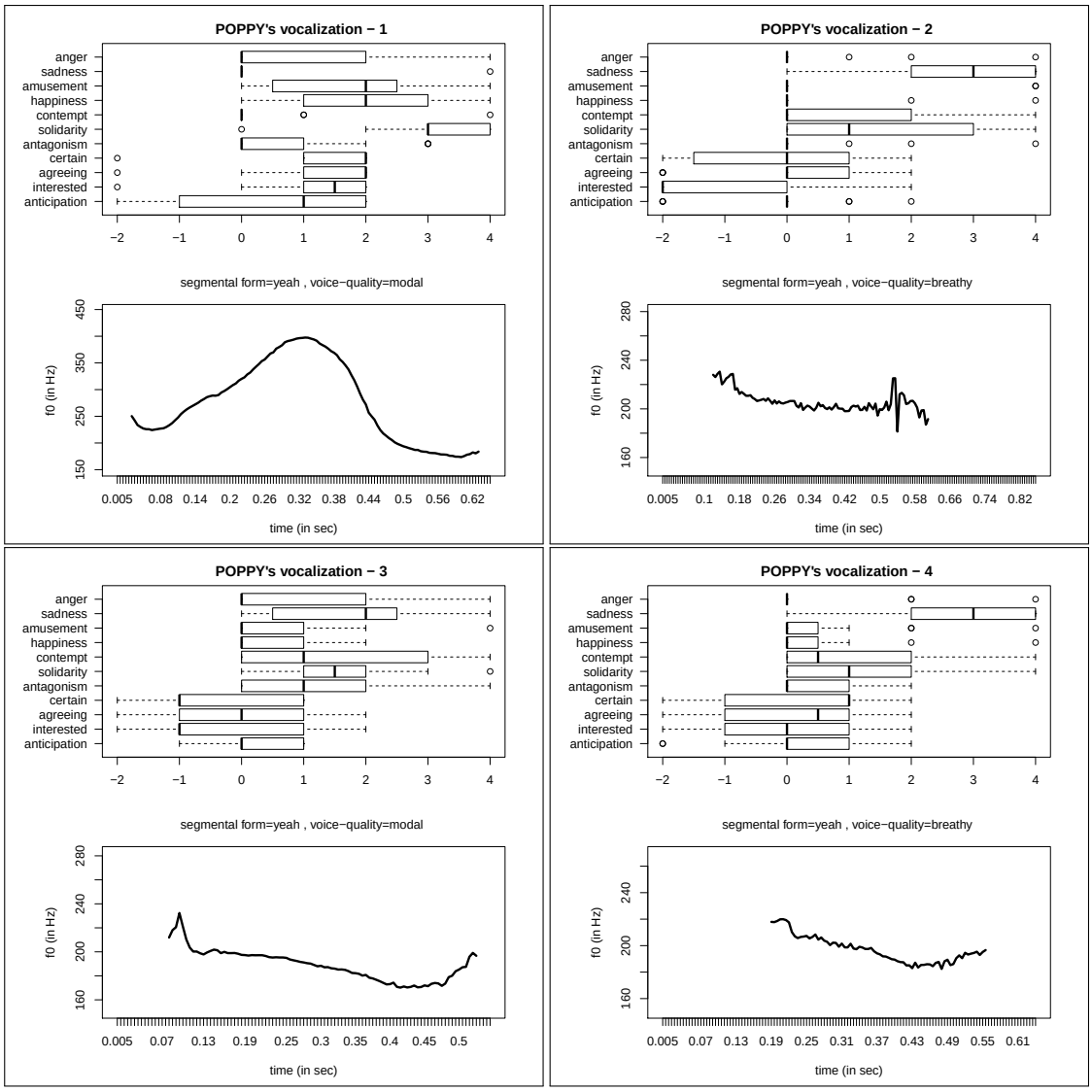


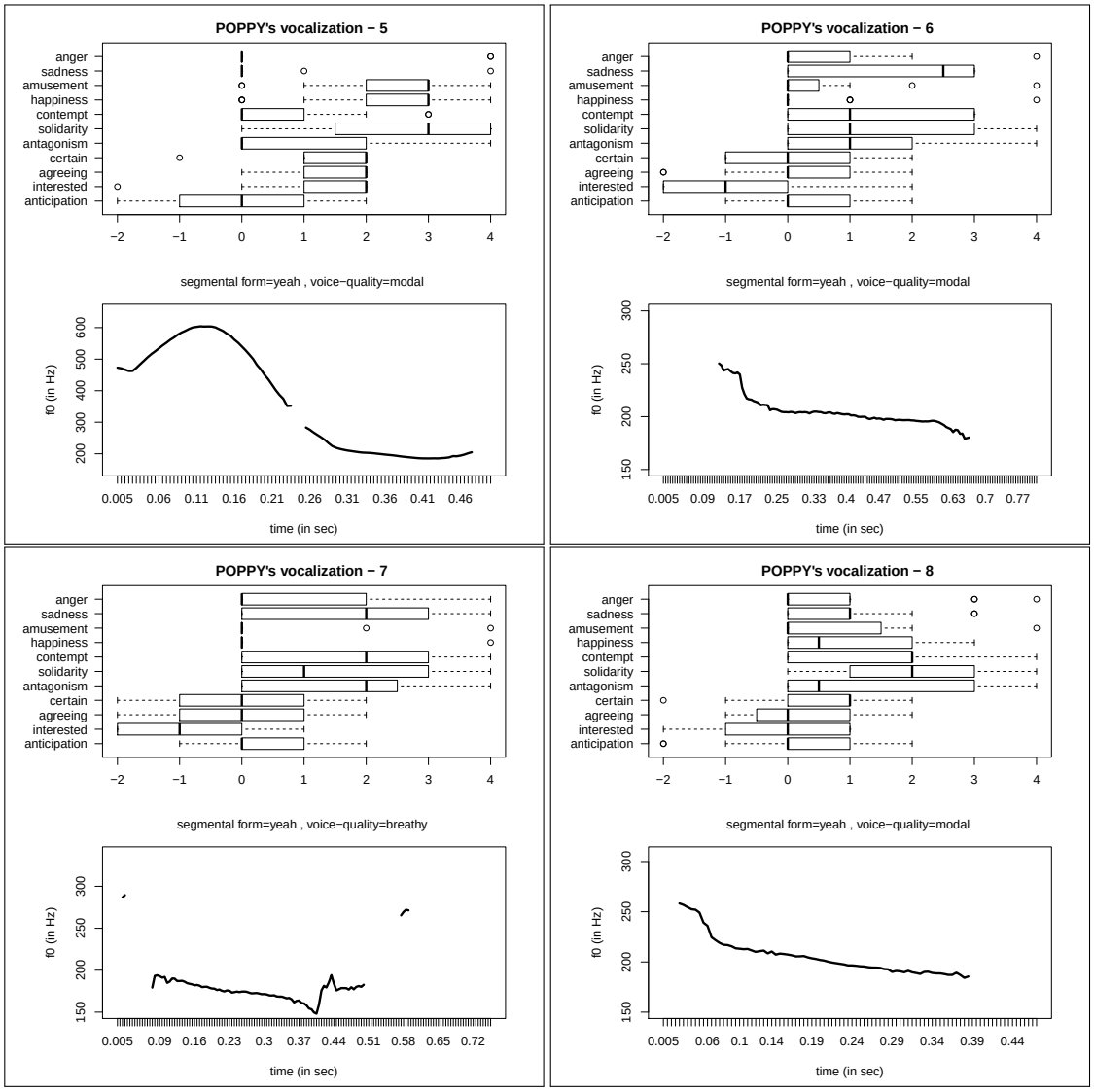
Appendix B

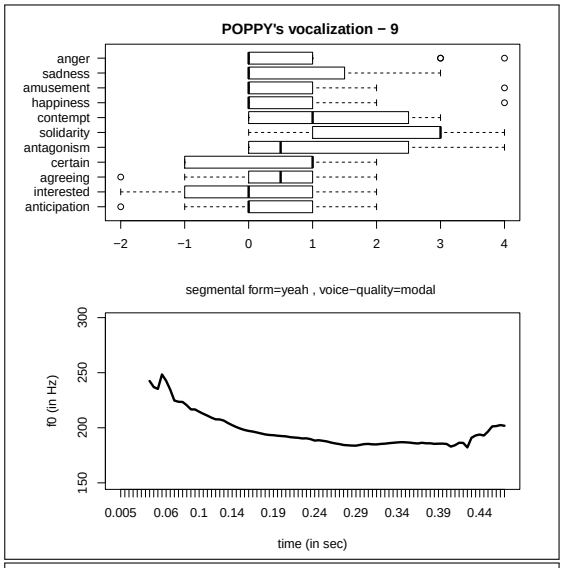
Annotation of British English vocalizations

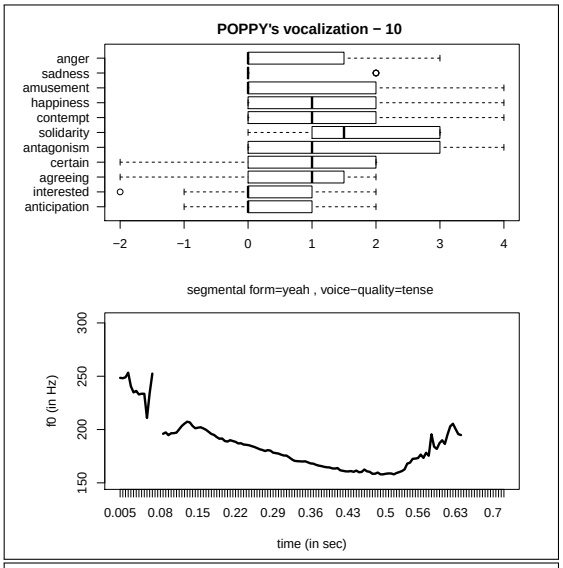
The annotation of selected stimuli is presented in this appendix. Each figure represents annotation of a vocalization, and it consists of: (i) meaning annotation of the vocalization in the form of a box plot; and (ii) annotation of behavior properties such as *segmental form, intonation contour*, and *voice quality*. Among 11 meaning scales used in meaning annotation, *anger, sadness, amusement, happiness, contempt, solidarity* and *antagonism* are unipolar (i.e. the scale values are in the range from 0 to 4); and remaining scales *(un)certain, (dis)agreeing, (un)interested*, and *(high/low)anticipation* are bipolar (i.e. the scale values are in the range from -2 to 2).

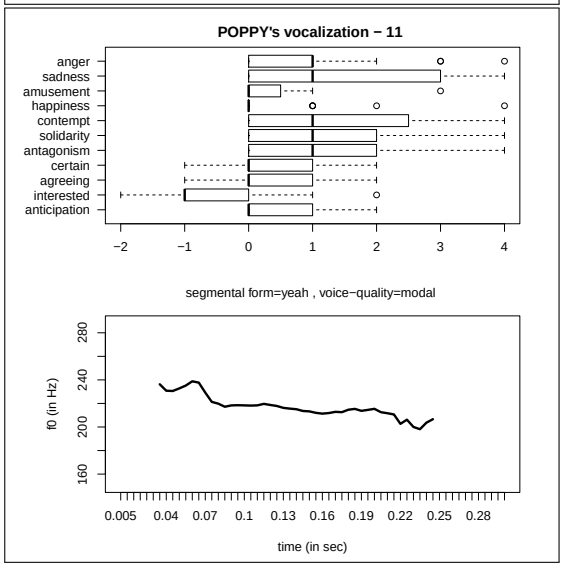
Poppy’s vocalizations

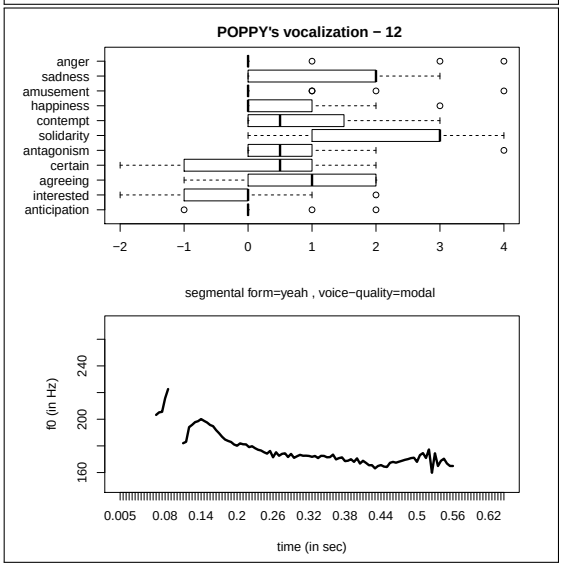


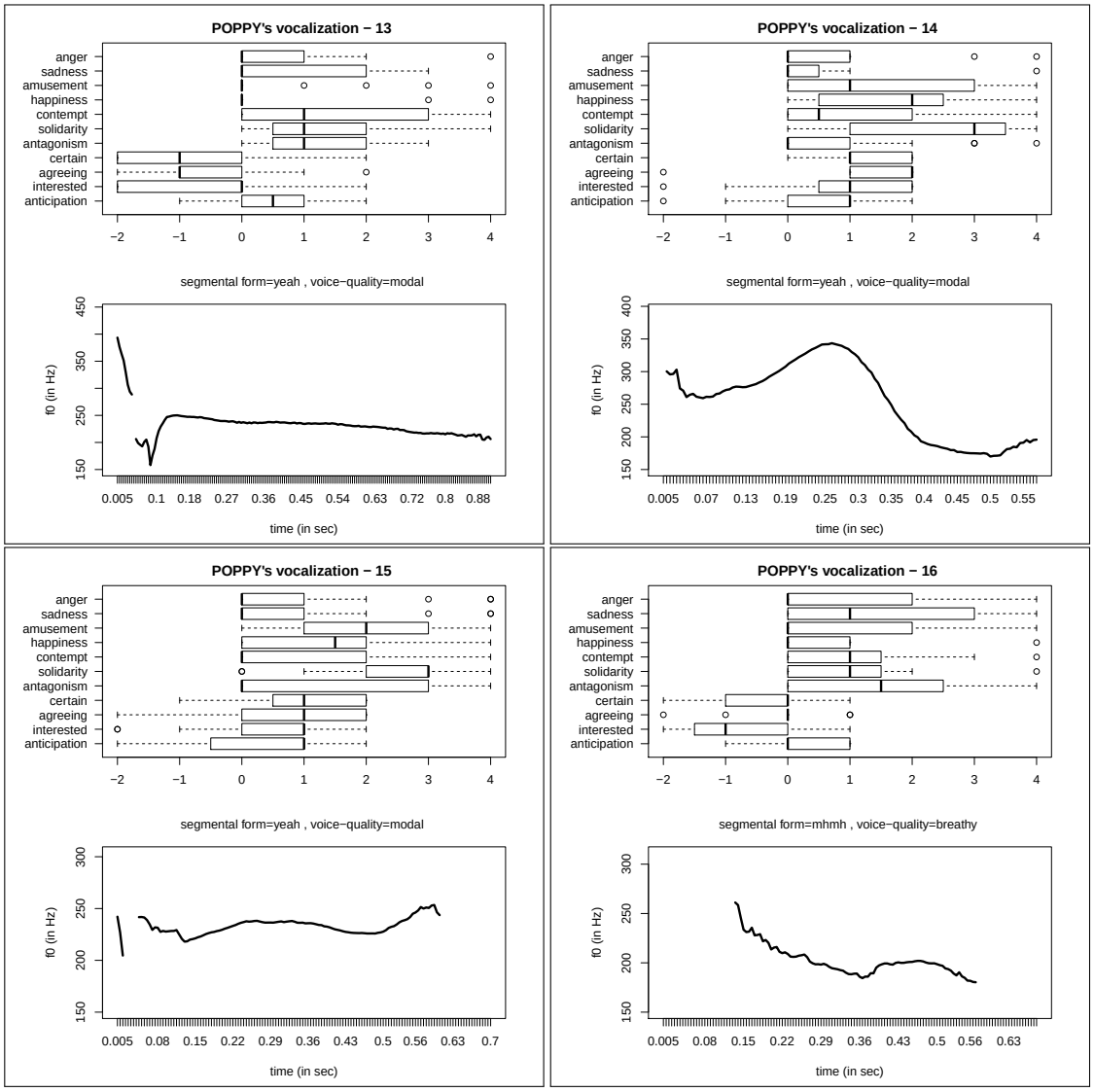

POPPY's vocalization − 5
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
200 300 400 500 600
0.005 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46
time (in sec)
POPPY's vocalization − 6
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250 300
0.005 0.09 0.17 0.25 0.33 0.4 0.47 0.55 0.63 0.7 0.77
time (in sec)
POPPY's vocalization − 7
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=breathy
f0 (in Hz)
150 200 250 300
0.005 0.09 0.16 0.23 0.3 0.37 0.44 0.51 0.58 0.65 0.72
time (in sec)
POPPY's vocalization − 8
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250 300
0.005 0.06 0.1 0.14 0.19 0.24 0.29 0.34 0.39 0.44
time (in sec)

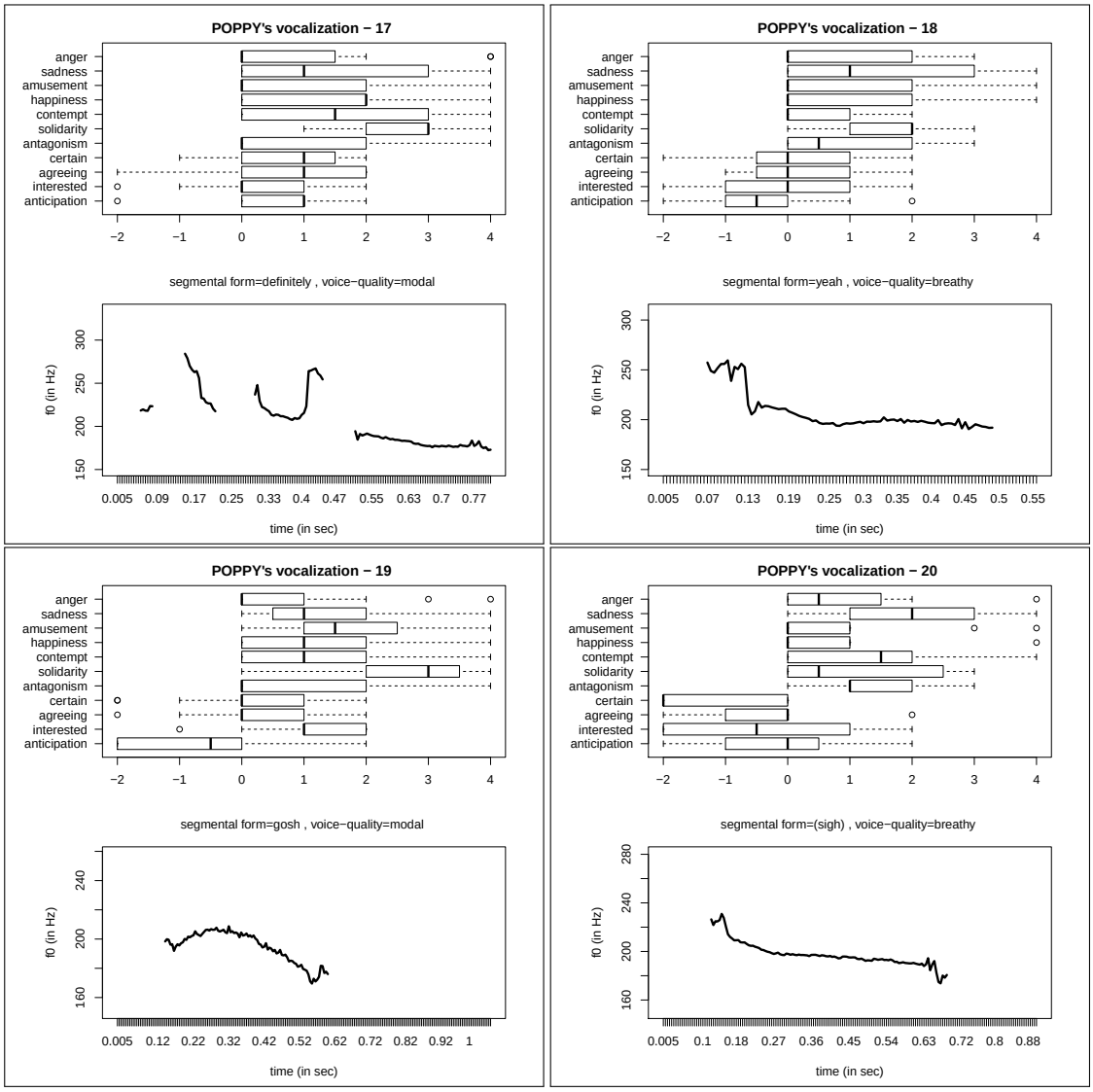

POPPY's vocalization − 9
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250 300
0.005 0.06 0.1 0.14 0.19 0.24 0.29 0.34 0.39 0.44
time (in sec)

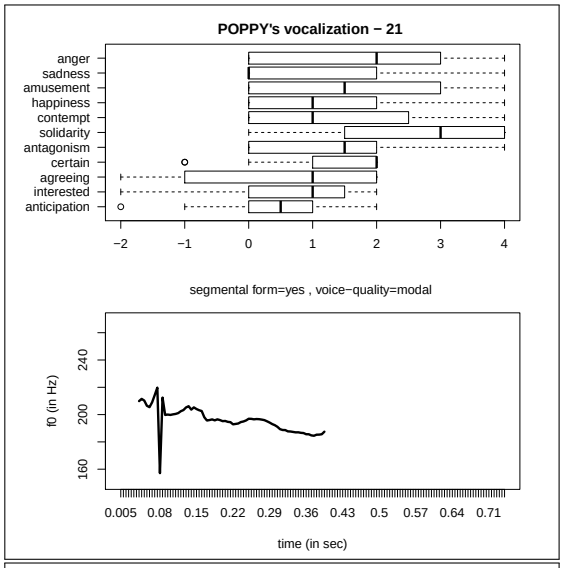

POPPY's vocalization − 10
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=tense
f0 (in Hz)
150 200 250 300
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63 0.7
time (in sec)

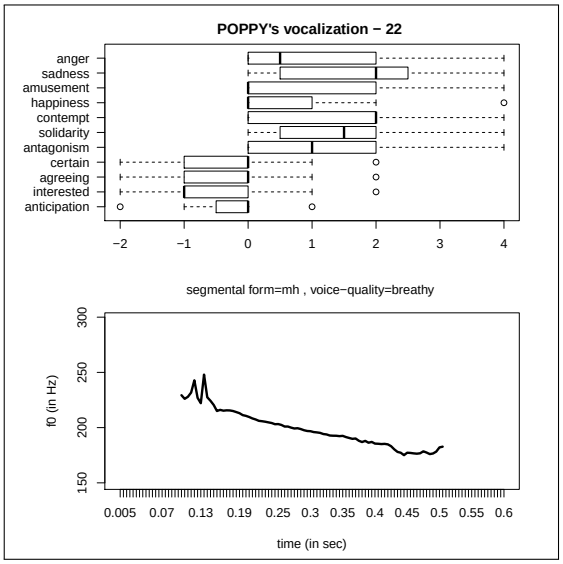

POPPY's vocalization − 11
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
160 200 240 280
0.005 0.04 0.07 0.1 0.13 0.16 0.19 0.22 0.25 0.28
time (in sec)

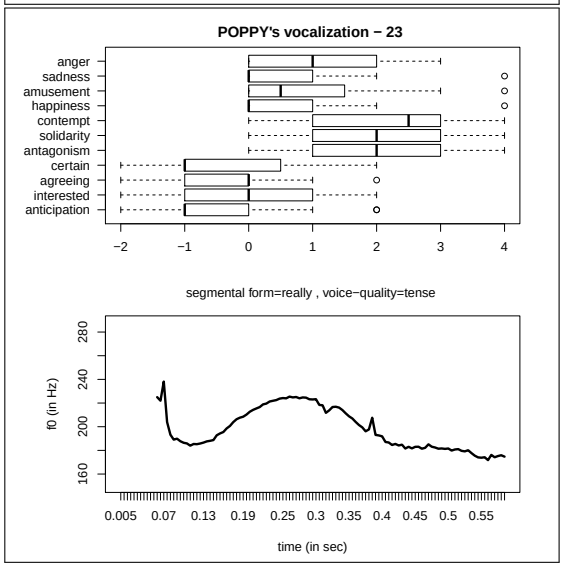

POPPY's vocalization − 12
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
160 200 240
0.005 0.08 0.14 0.2 0.26 0.32 0.38 0.44 0.5 0.56 0.62
time (in sec)


POPPY's vocalization − 13
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 250 350 450
0.005 0.1 0.18 0.27 0.36 0.45 0.54 0.63 0.72 0.8 0.88
time (in sec)
POPPY's vocalization − 14
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250 300 350 400
0.005 0.07 0.13 0.19 0.25 0.3 0.35 0.4 0.45 0.5 0.55
time (in sec)
POPPY's vocalization − 15
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250 300
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63 0.7
time (in sec)
POPPY's vocalization − 16
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=mhmh , voice−quality=breathy
f0 (in Hz)
150 200 250 300
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63
time (in sec)

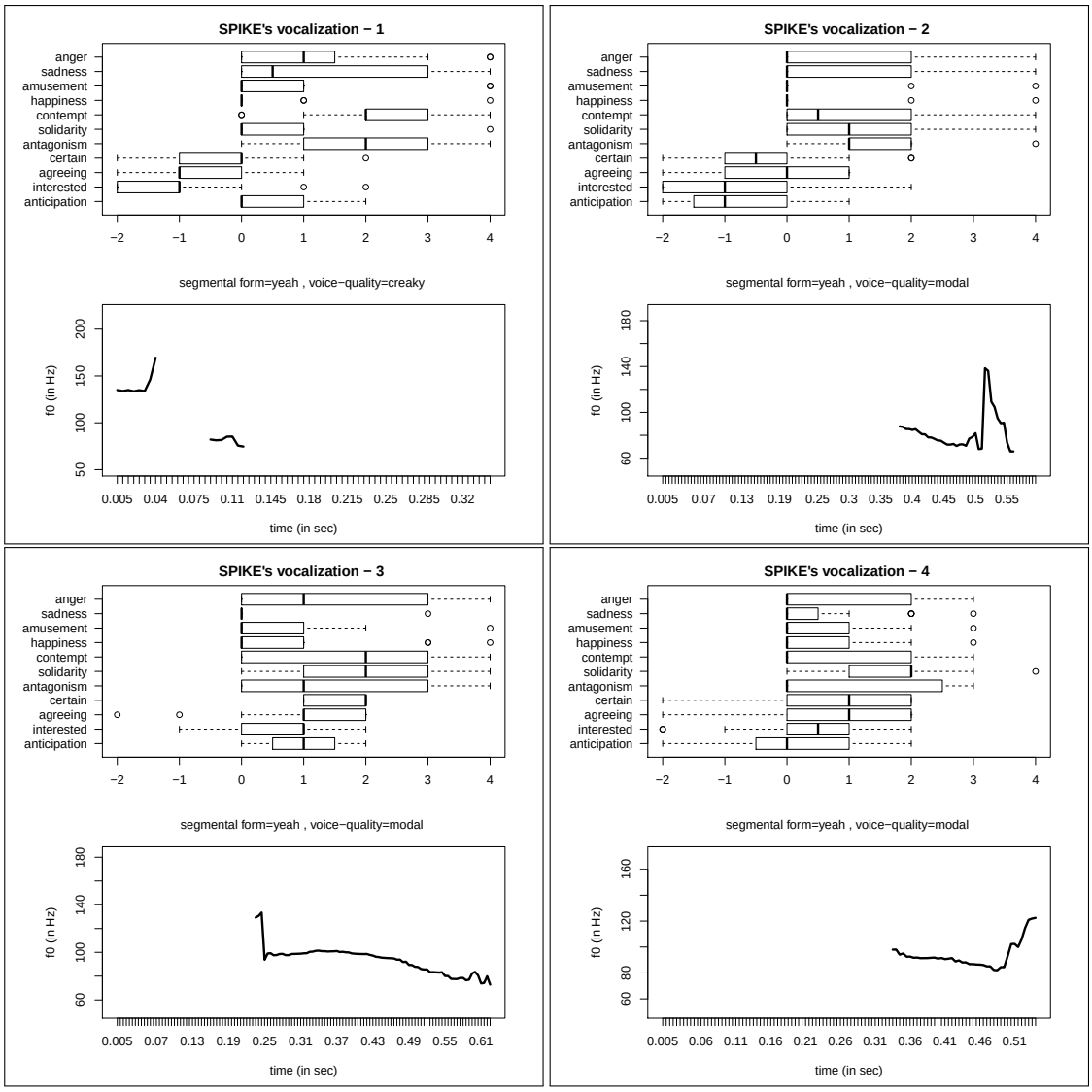

POPPY's vocalization − 17
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=definitely , voice−quality=modal
f0 (in Hz)
150 200 250 300
0.005 0.09 0.17 0.25 0.33 0.4 0.47 0.55 0.63 0.7 0.77
time (in sec)
POPPY's vocalization − 18
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=breathy
f0 (in Hz)
150 200 250 300
0.005 0.07 0.13 0.19 0.25 0.3 0.35 0.4 0.45 0.5 0.55
time (in sec)
POPPY's vocalization − 19
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=gosh , voice−quality=modal
f0 (in Hz)
160 200 240
0.005 0.12 0.22 0.32 0.42 0.52 0.62 0.72 0.82 0.92 1
time (in sec)
POPPY's vocalization − 20
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=(sigh) , voice−quality=breathy
f0 (in Hz)
160 200 240 280
0.005 0.1 0.18 0.27 0.36 0.45 0.54 0.63 0.72 0.8 0.88
time (in sec)

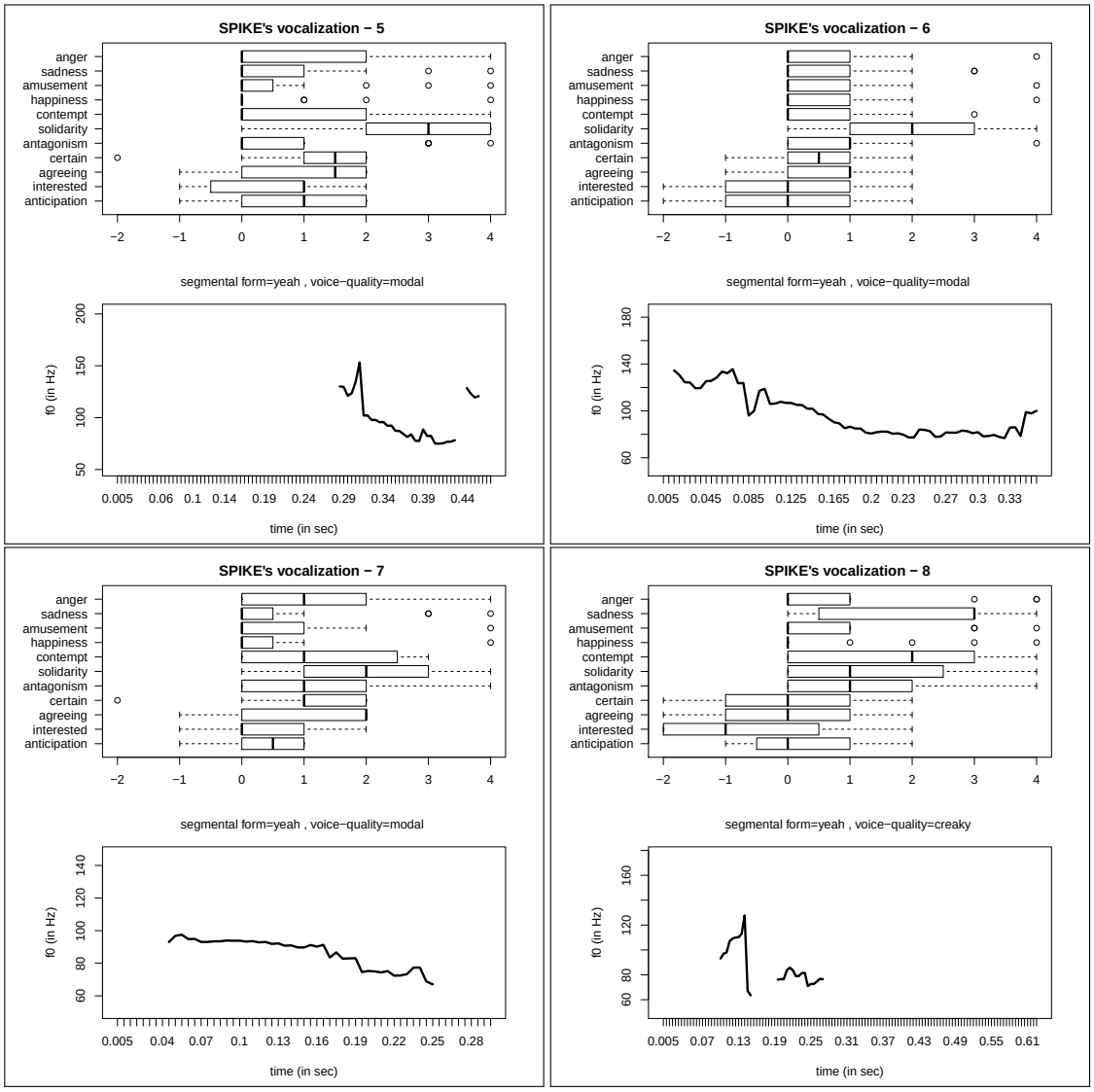

POPPY's vocalization − 21
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yes , voice−quality=modal
f0 (in Hz)
160 200 240
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.57 0.64 0.71
time (in sec)

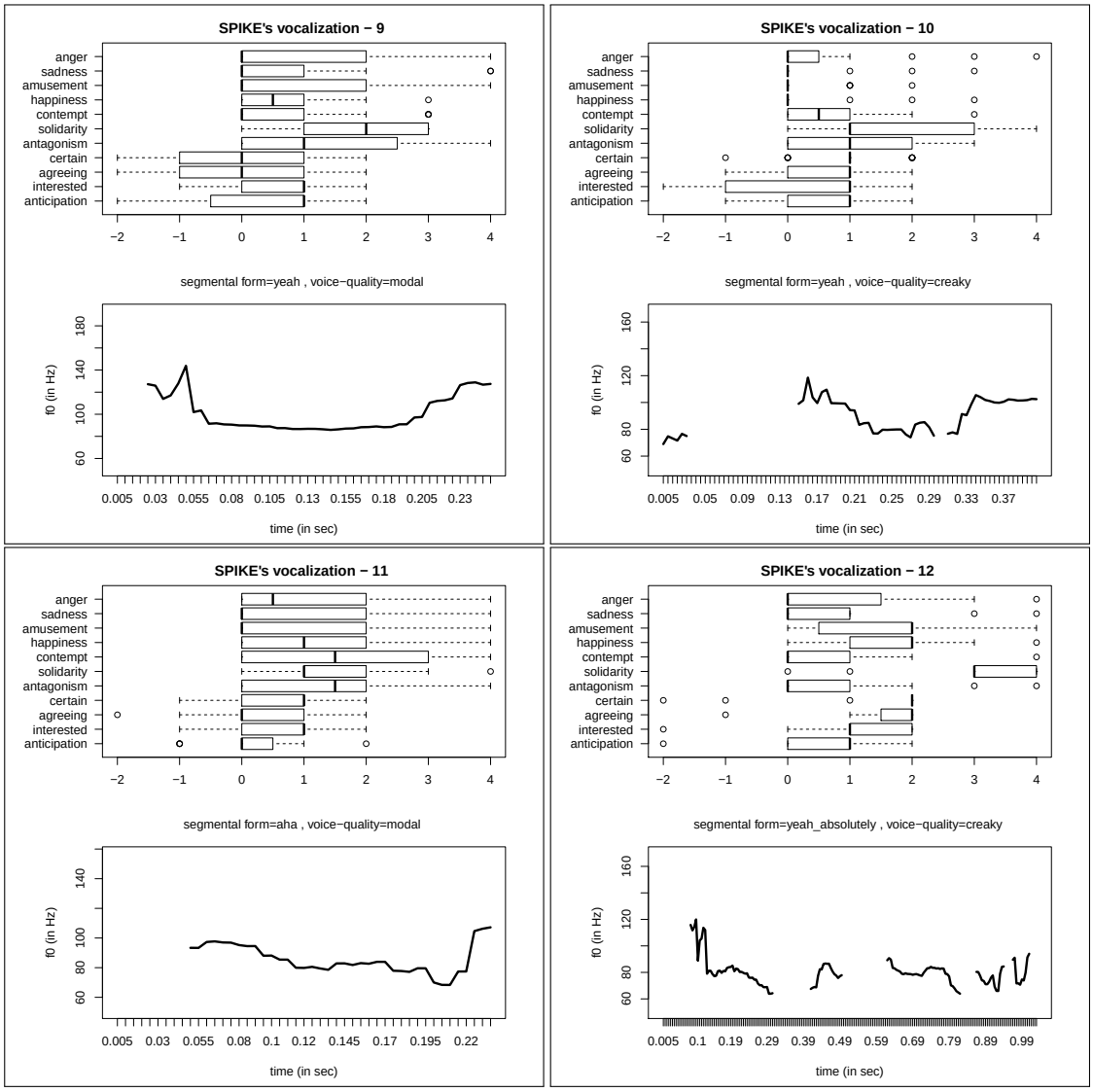

POPPY's vocalization − 22
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=mh , voice−quality=breathy
f0 (in Hz)
150 200 250 300
0.005 0.07 0.13 0.19 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
time (in sec)

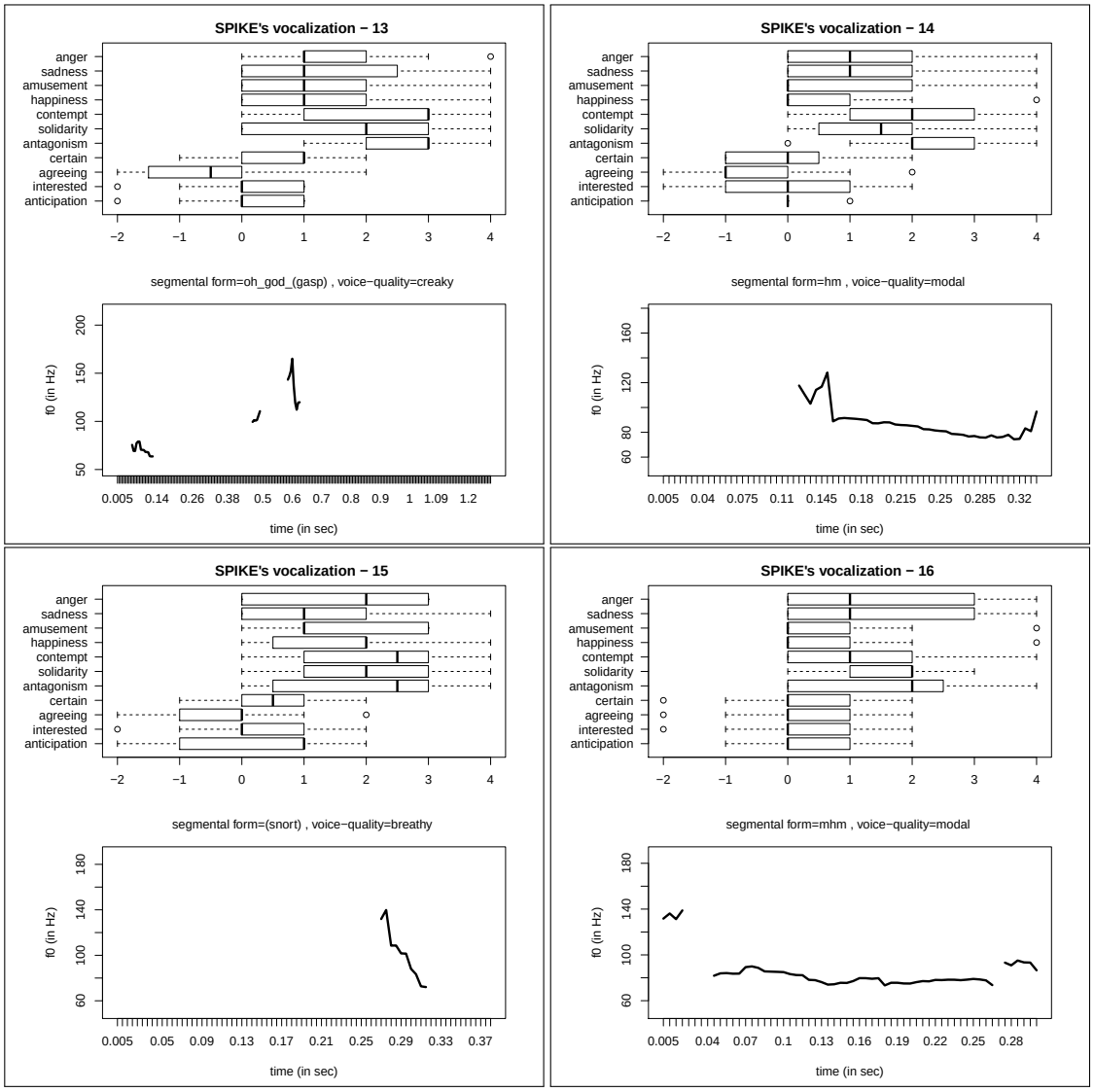

POPPY's vocalization − 23
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=really , voice−quality=tense
f0 (in Hz)
160 200 240 280
0.005 0.07 0.13 0.19 0.25 0.3 0.35 0.4 0.45 0.5 0.55
time (in sec)

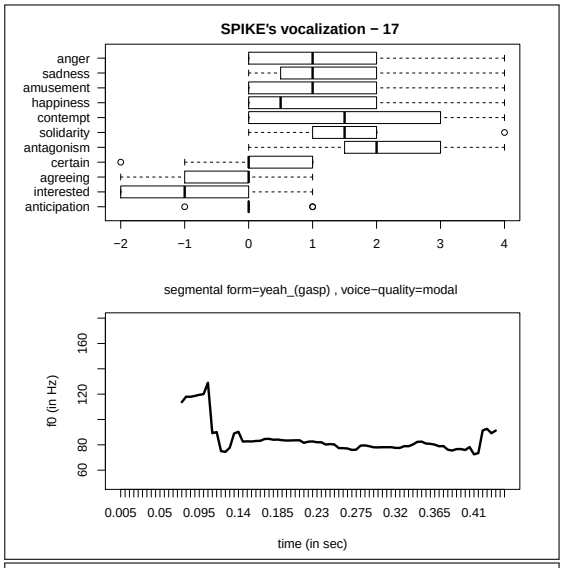
Spike's vocalizations

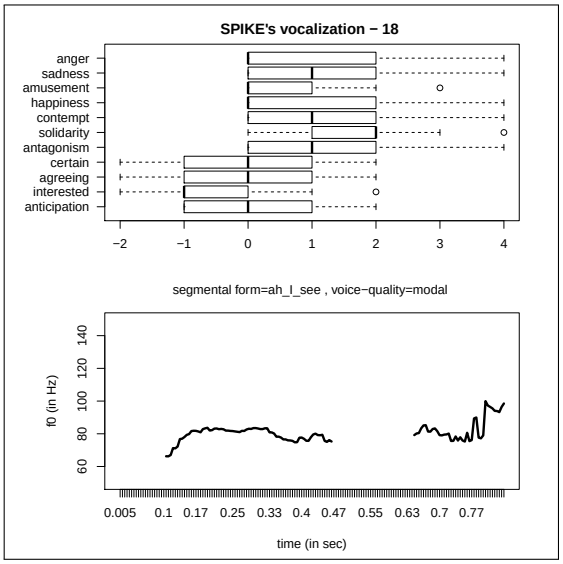


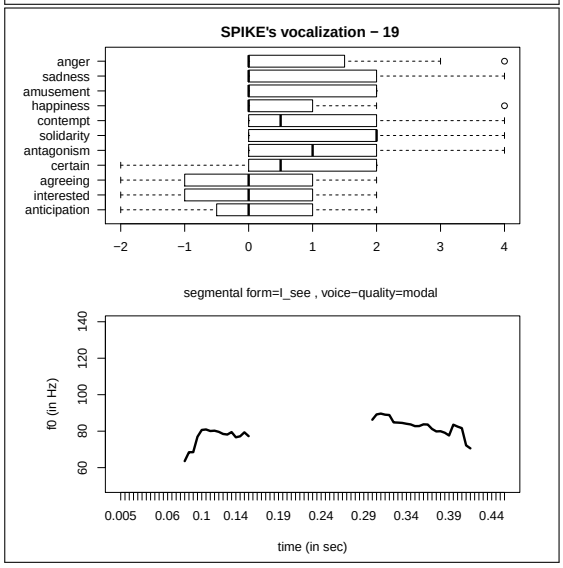

SPIKE's vocalization − 5
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
50 100 150 200
0.005 0.06 0.1 0.14 0.19 0.24 0.29 0.34 0.39 0.44
time (in sec)
SPIKE's vocalization − 6
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
60 100 140 180
0.005 0.045 0.085 0.125 0.165 0.2 0.23 0.27 0.3 0.33
time (in sec)
SPIKE's vocalization − 7
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
60 80 100 120 140
0.005 0.04 0.07 0.1 0.13 0.16 0.19 0.22 0.25 0.28
time (in sec)
SPIKE's vocalization − 8
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=creaky
f0 (in Hz)
60 80 120 160
0.005 0.07 0.13 0.19 0.25 0.31 0.37 0.43 0.49 0.55 0.61
time (in sec)


SPIKE's vocalization − 9
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
60 100 140 180
0.005 0.03 0.055 0.08 0.105 0.13 0.155 0.18 0.205 0.23
time (in sec)
SPIKE's vocalization − 10
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=creaky
f0 (in Hz)
60 80 120 160
0.005 0.05 0.09 0.13 0.17 0.21 0.25 0.29 0.33 0.37
time (in sec)
SPIKE's vocalization − 11
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=aha , voice−quality=modal
f0 (in Hz)
60 80 100 140
0.005 0.03 0.055 0.08 0.1 0.12 0.145 0.17 0.195 0.22
time (in sec)
SPIKE's vocalization − 12
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah_absolutely , voice−quality=creaky
f0 (in Hz)
60 80 120 160
0.005 0.1 0.19 0.29 0.39 0.49 0.59 0.69 0.79 0.89 0.99
time (in sec)

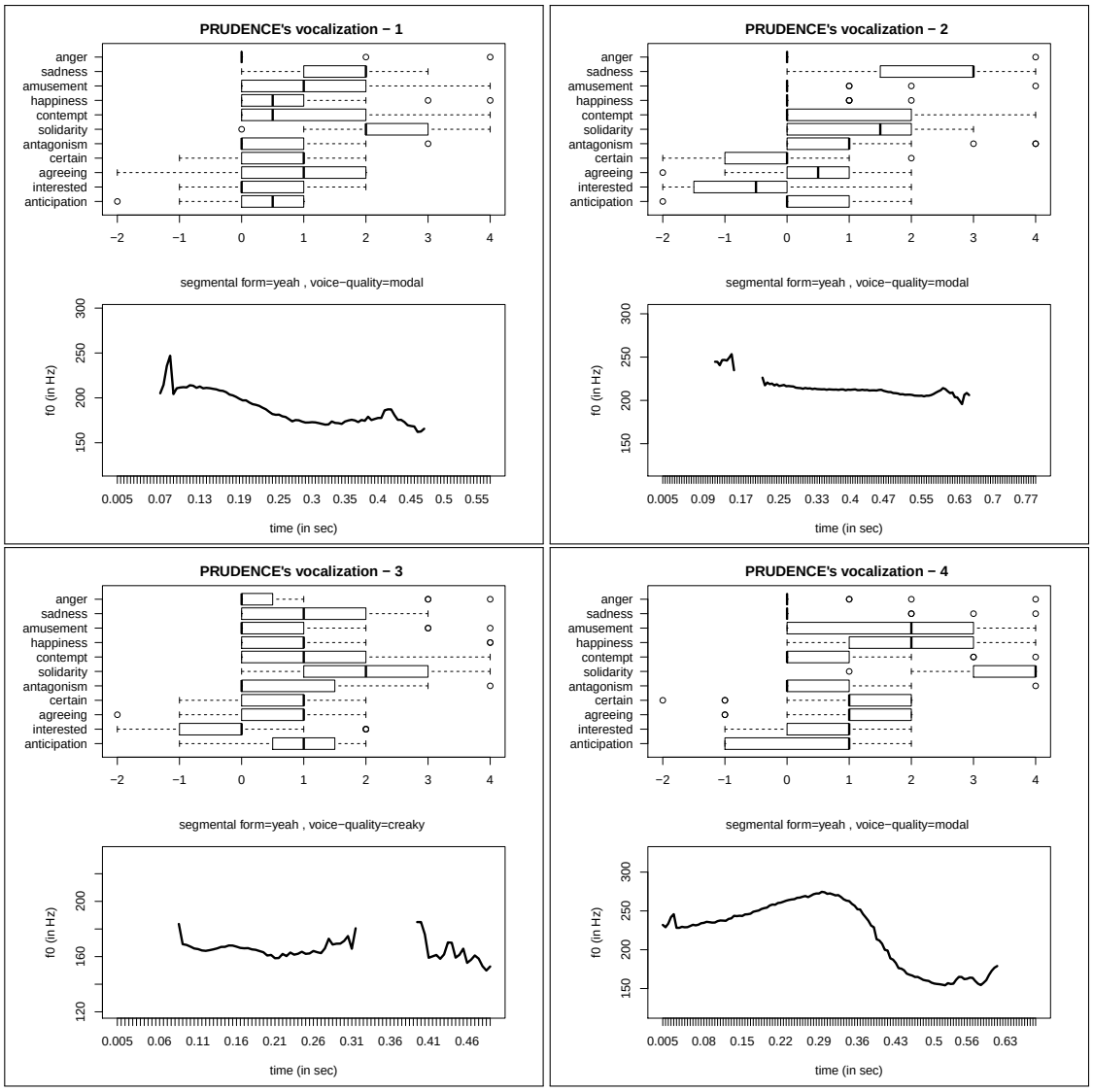

SPIKE's vocalization − 13
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=oh_god_(gasp) , voice−quality=creaky
f0 (in Hz)
50 100 150 200
0.005 0.14 0.26 0.38 0.5 0.6 0.7 0.8 0.9 1 1.09 1.2
time (in sec)
SPIKE's vocalization − 14
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=hm , voice−quality=modal
f0 (in Hz)
60 80 120 160
0.005 0.04 0.075 0.11 0.145 0.18 0.215 0.25 0.285 0.32
time (in sec)
SPIKE's vocalization − 15
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=(snort) , voice−quality=breathy
f0 (in Hz)
60 100 140 180
0.005 0.05 0.09 0.13 0.17 0.21 0.25 0.29 0.33 0.37
time (in sec)
SPIKE's vocalization − 16
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=mhm , voice−quality=modal
f0 (in Hz)
60 100 140 180
0.005 0.04 0.07 0.1 0.13 0.16 0.19 0.22 0.25 0.28
time (in sec)

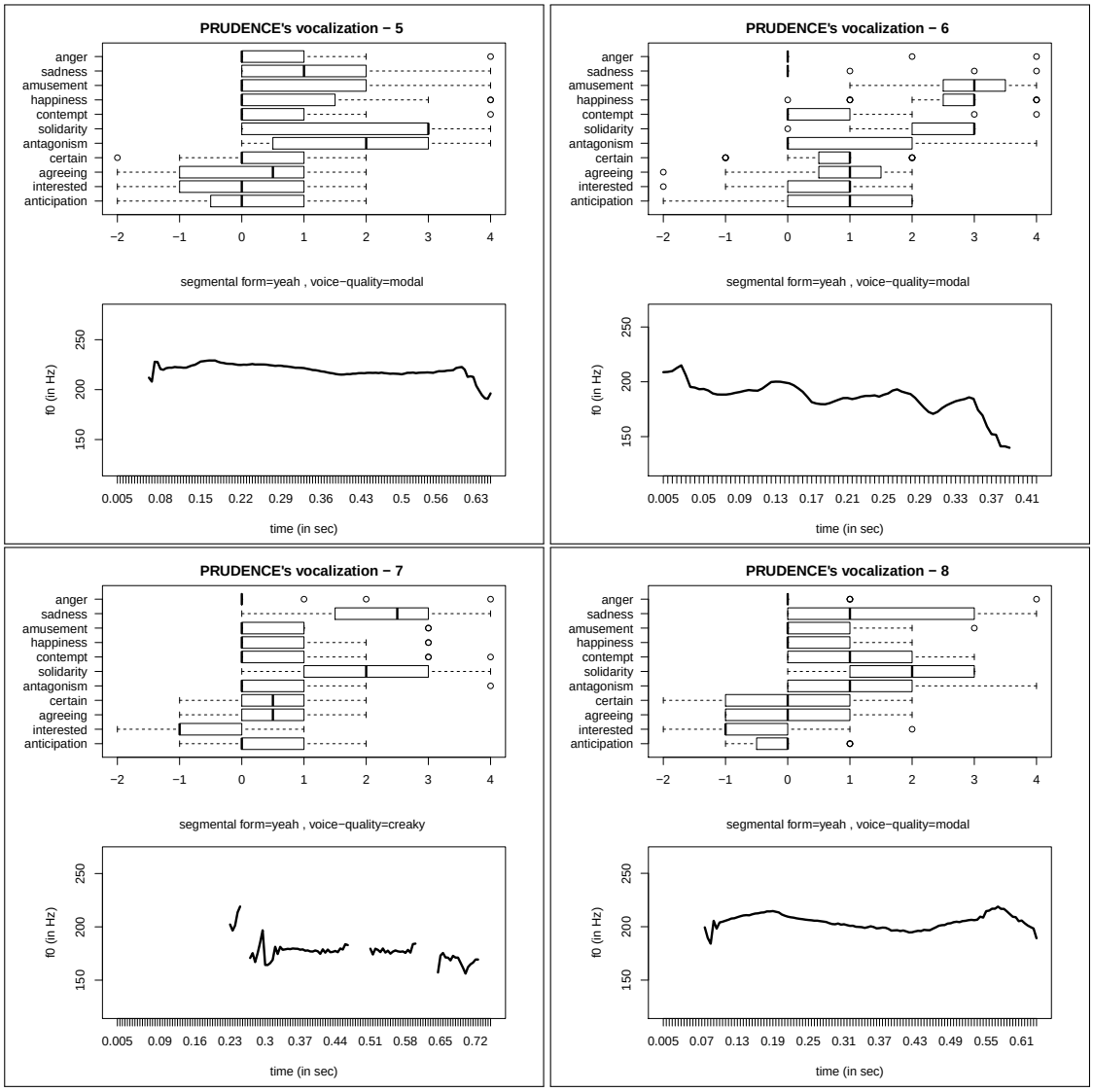

SPIKE's vocalization − 17
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah_(gasp) , voice−quality=modal
f0 (in Hz)
60 80 120 160
0.005 0.05 0.095 0.14 0.185 0.23 0.275 0.32 0.365 0.41
time (in sec)

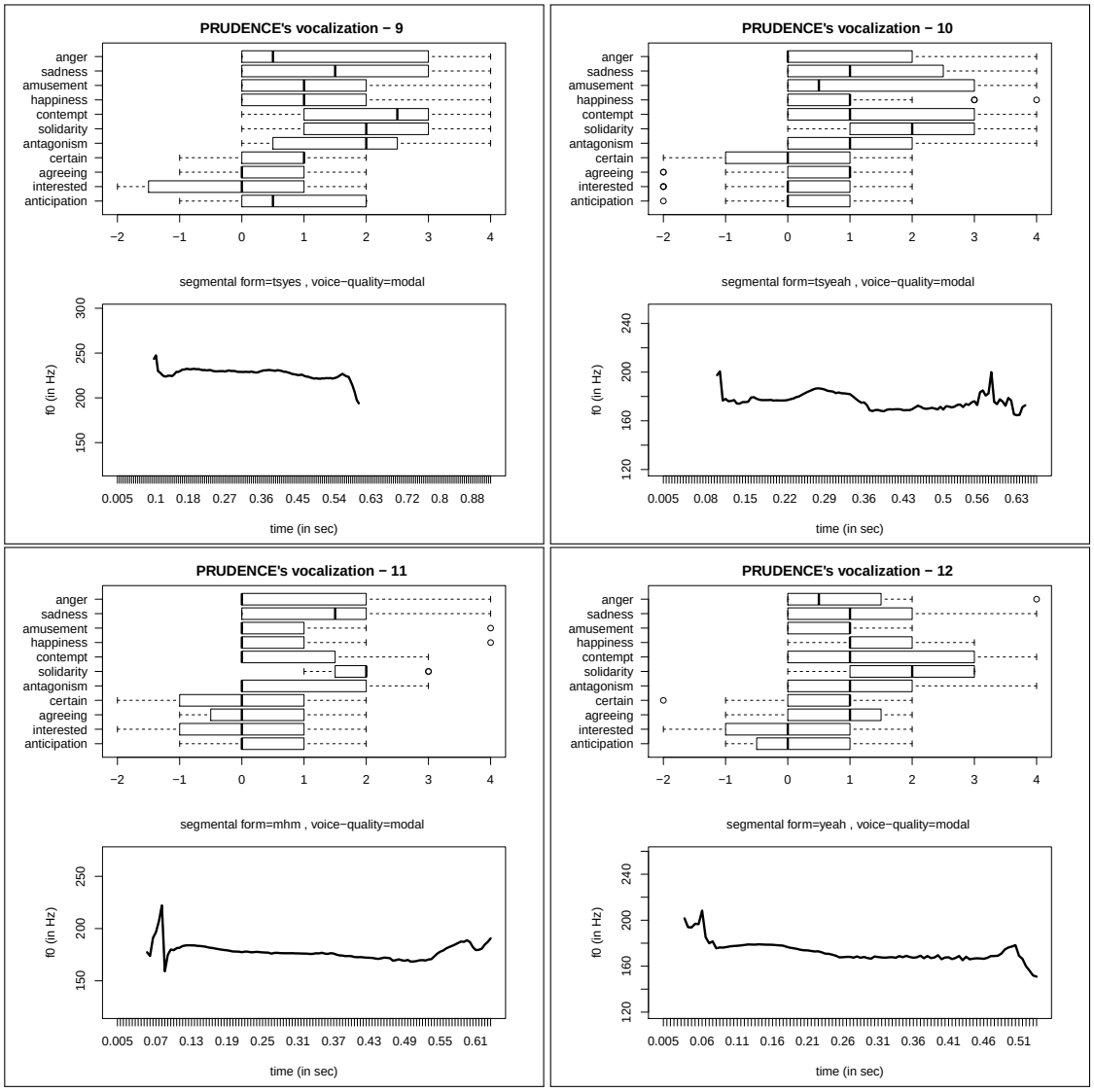

SPIKE's vocalization − 18
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=ah_I_see , voice−quality=modal
f0 (in Hz)
60 80 100 120 140
0.005 0.1 0.17 0.25 0.33 0.4 0.47 0.55 0.63 0.7 0.77
time (in sec)

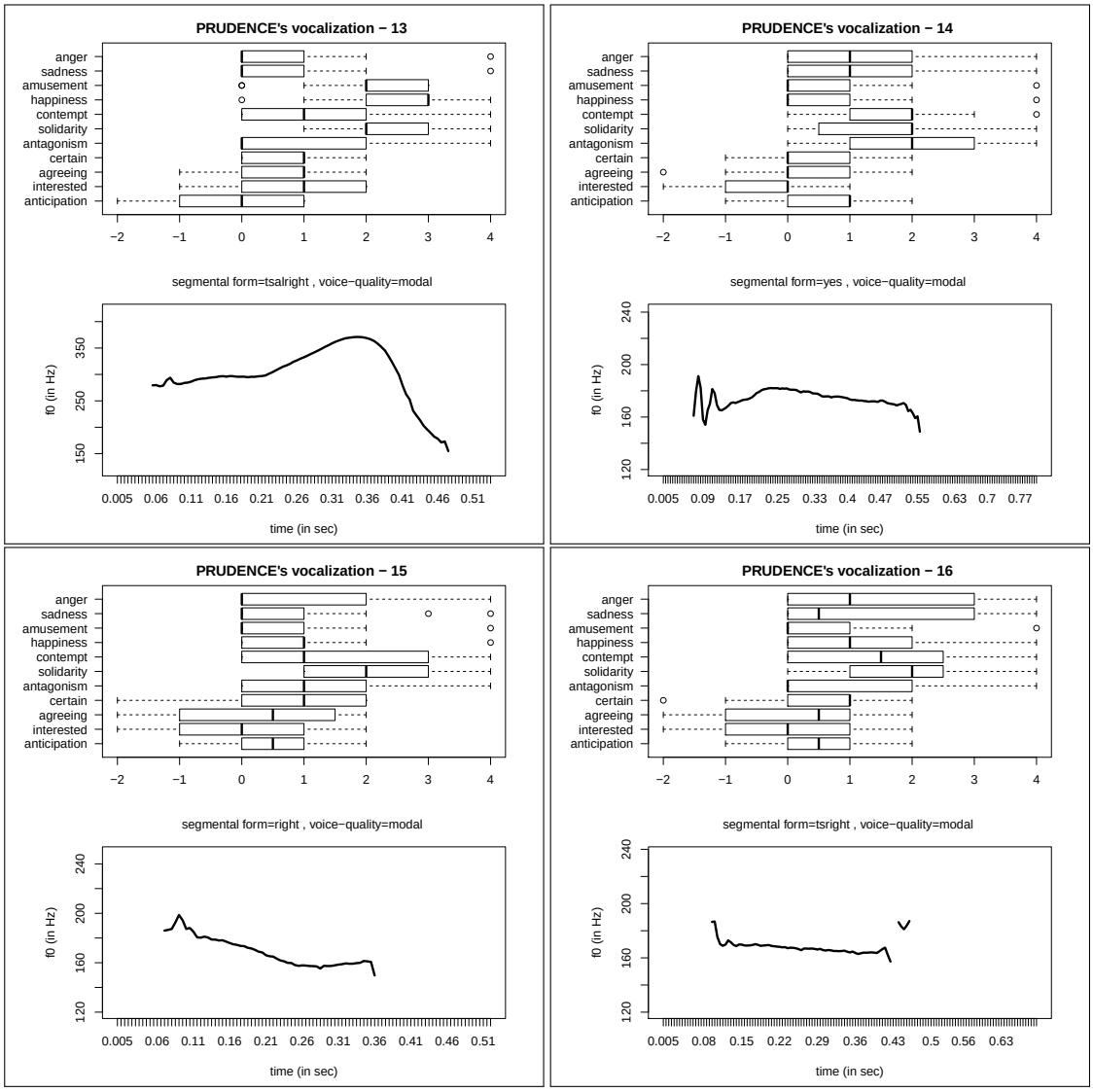

SPIKE's vocalization − 19
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=I_see , voice−quality=modal
f0 (in Hz)
60 80 100 120 140
0.005 0.06 0.1 0.14 0.19 0.24 0.29 0.34 0.39 0.44
time (in sec)

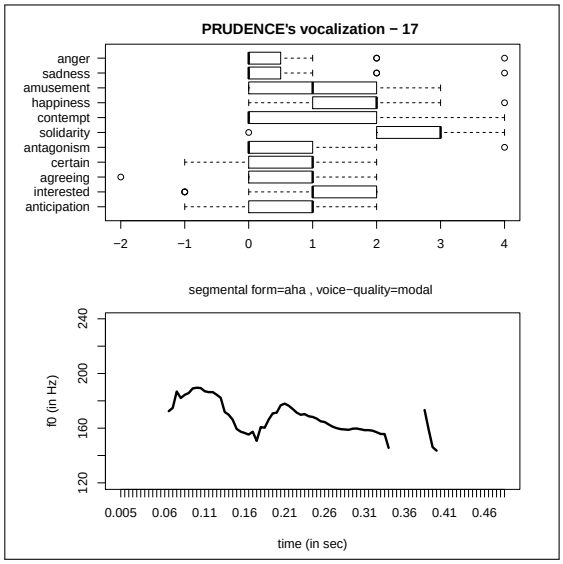
Prudence’s vocalizations

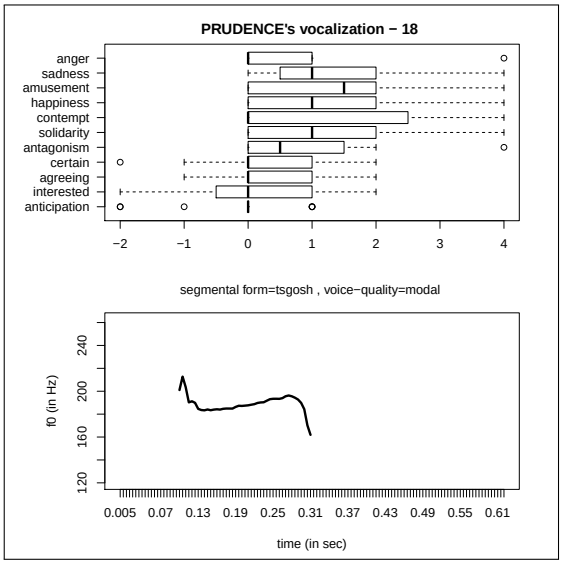



PRUDENCE's vocalization − 5
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63
time (in sec)
PRUDENCE's vocalization − 6
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250
0.005 0.05 0.09 0.13 0.17 0.21 0.25 0.29 0.33 0.37 0.41
time (in sec)
PRUDENCE's vocalization − 7
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=creaky
f0 (in Hz)
150 200 250
0.005 0.09 0.16 0.23 0.3 0.37 0.44 0.51 0.58 0.65 0.72
time (in sec)
PRUDENCE's vocalization − 8
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
150 200 250
0.005 0.07 0.13 0.19 0.25 0.31 0.37 0.43 0.49 0.55 0.61
time (in sec)

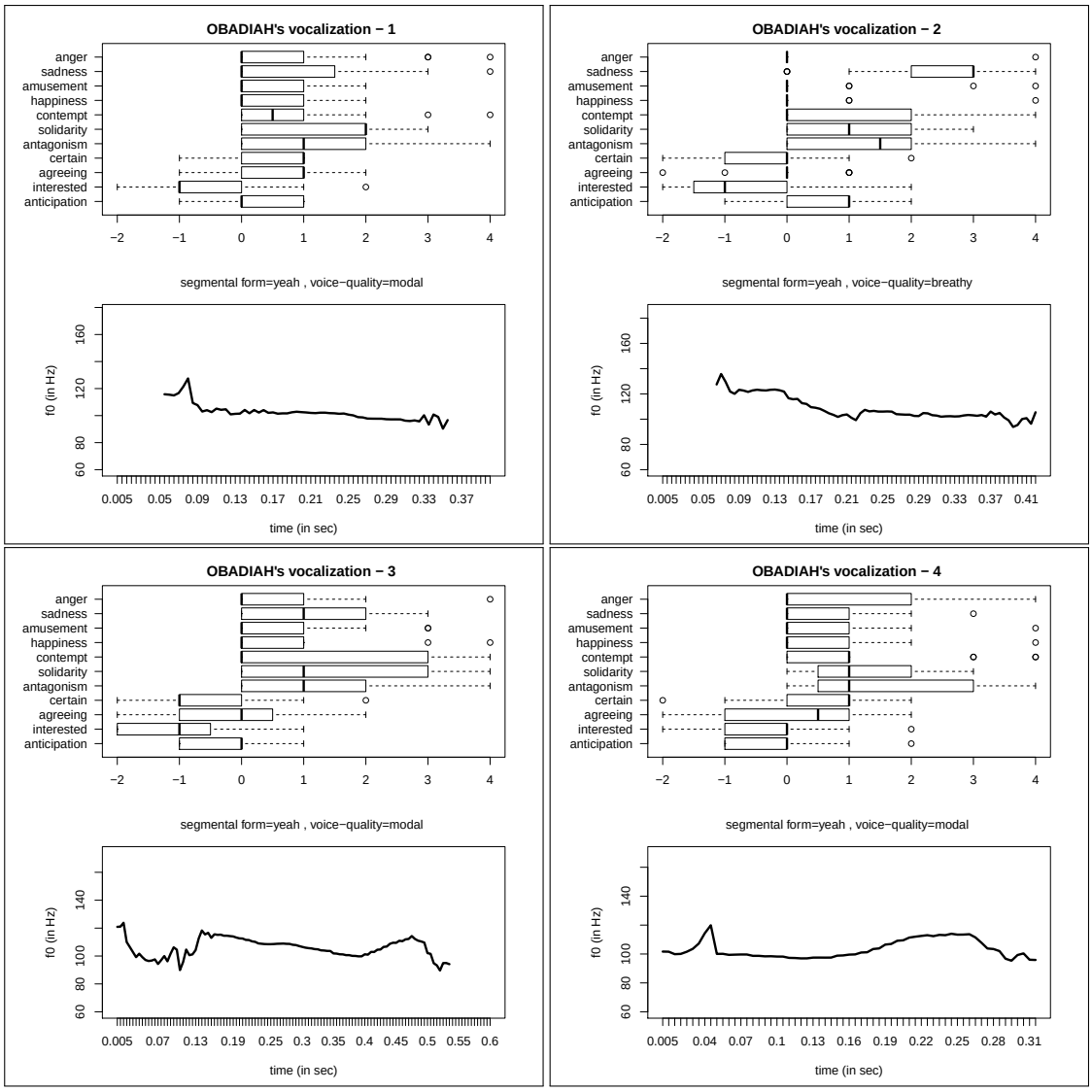

PRUDENCE's vocalization − 9
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=tsyes , voice−quality=modal
f0 (in Hz)
150 200 250 300
0.005 0.1 0.18 0.27 0.36 0.45 0.54 0.63 0.72 0.8 0.88
time (in sec)
PRUDENCE's vocalization − 10
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=tsyeah , voice−quality=modal
f0 (in Hz)
120 160 200 240
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63
time (in sec)
PRUDENCE's vocalization − 11
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=mhm , voice−quality=modal
f0 (in Hz)
150 200 250
0.005 0.07 0.13 0.19 0.25 0.31 0.37 0.43 0.49 0.55 0.61
time (in sec)
PRUDENCE's vocalization − 12
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
120 160 200 240
0.005 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46 0.51
time (in sec)

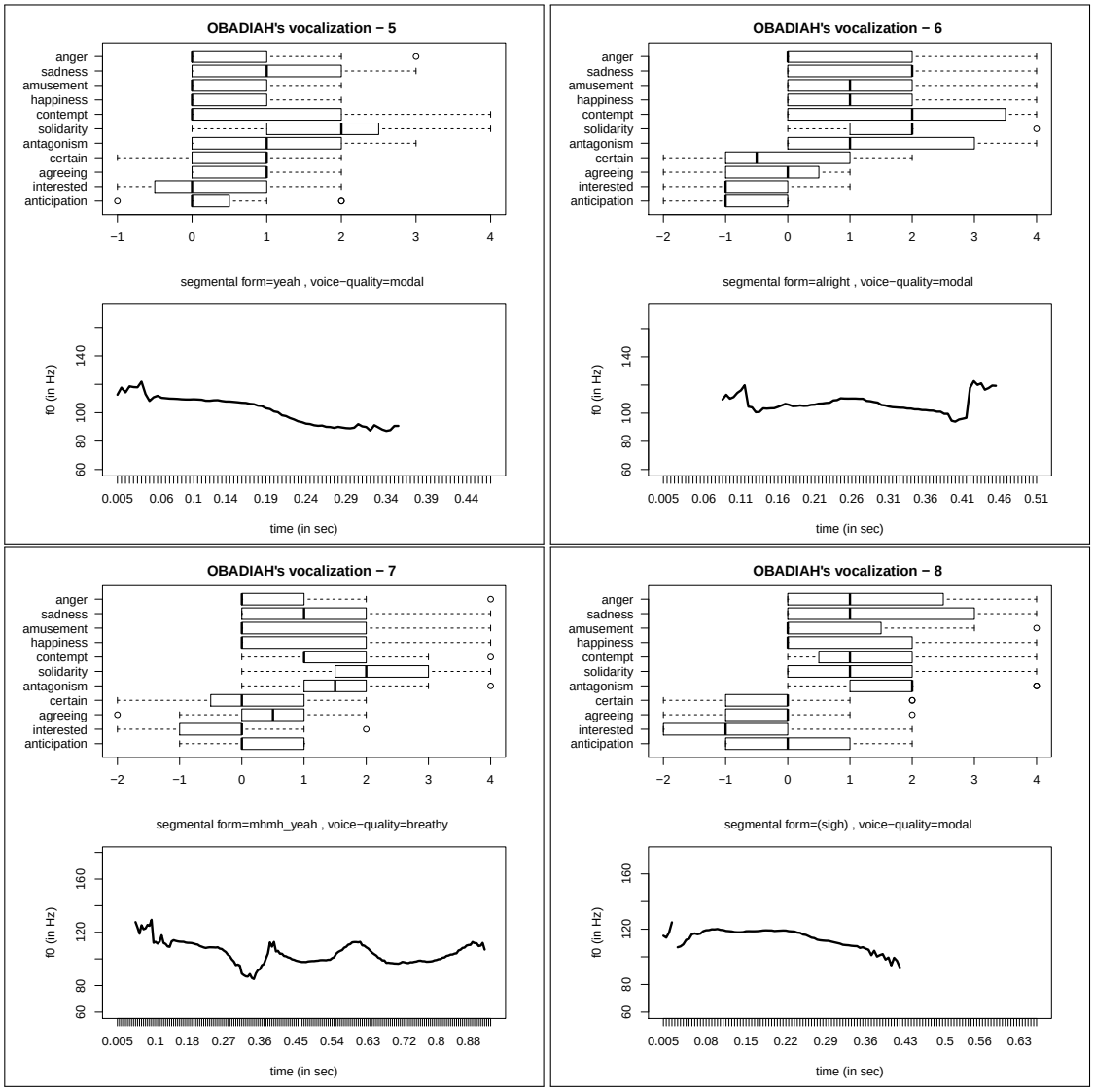

PRUDENCE's vocalization − 13
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=tsalright , voice−quality=modal
f0 (in Hz)
150 250 350
0.005 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46 0.51
time (in sec)
PRUDENCE's vocalization − 14
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yes , voice−quality=modal
f0 (in Hz)
120 160 200 240
0.005 0.09 0.17 0.25 0.33 0.4 0.47 0.55 0.63 0.7 0.77
time (in sec)
PRUDENCE's vocalization − 15
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=right , voice−quality=modal
f0 (in Hz)
120 160 200 240
0.005 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46 0.51
time (in sec)
PRUDENCE's vocalization − 16
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=tsright , voice−quality=modal
f0 (in Hz)
120 160 200 240
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63
time (in sec)

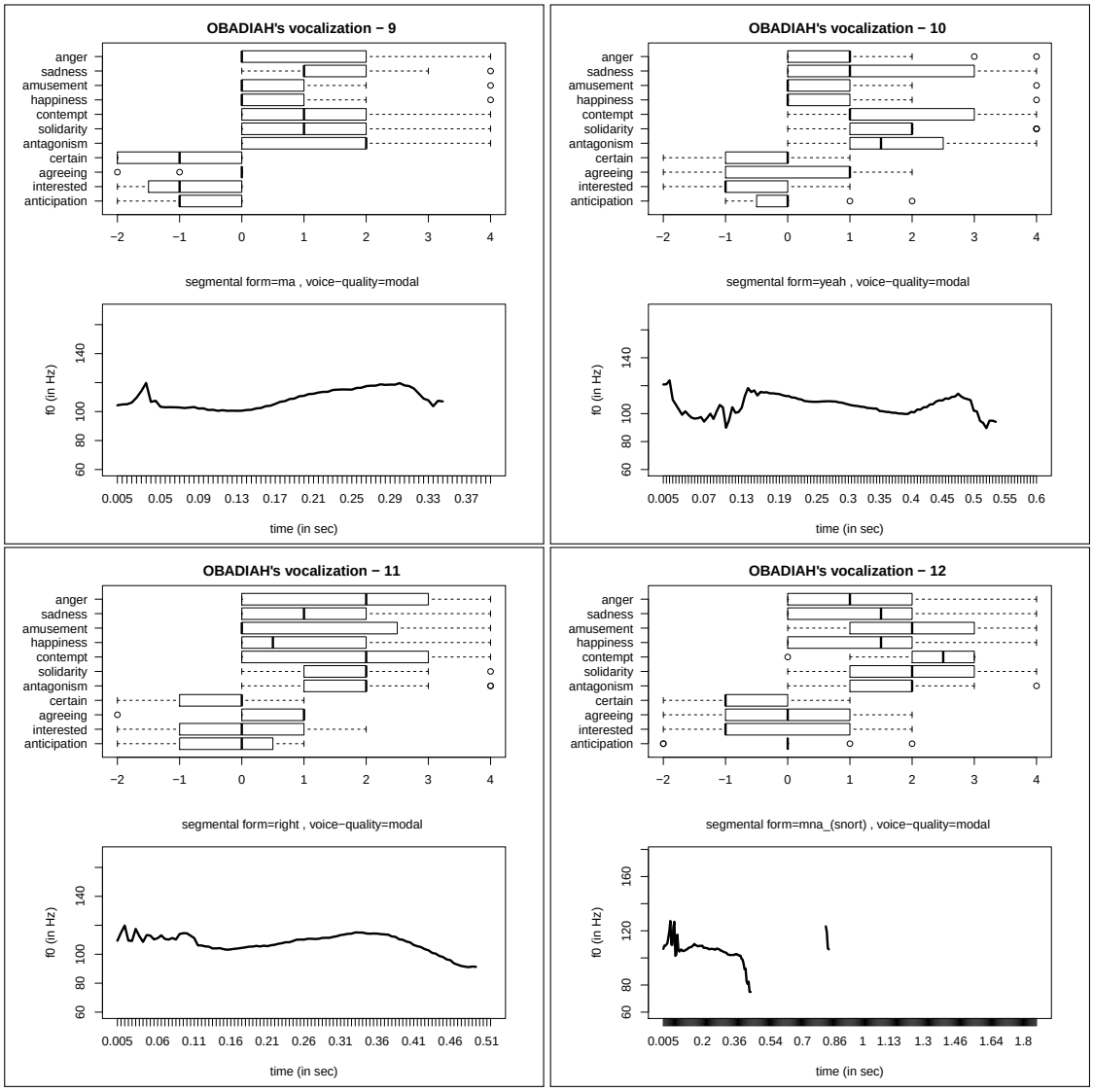

PRUDENCE's vocalization − 17
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
segmental form=aha , voice−quality=modal
f0 (in Hz)
120
160
200
240
0.005
0.06
0.11
0.16
0.21
0.26
0.31
0.36
0.41
0.46
time (in sec)

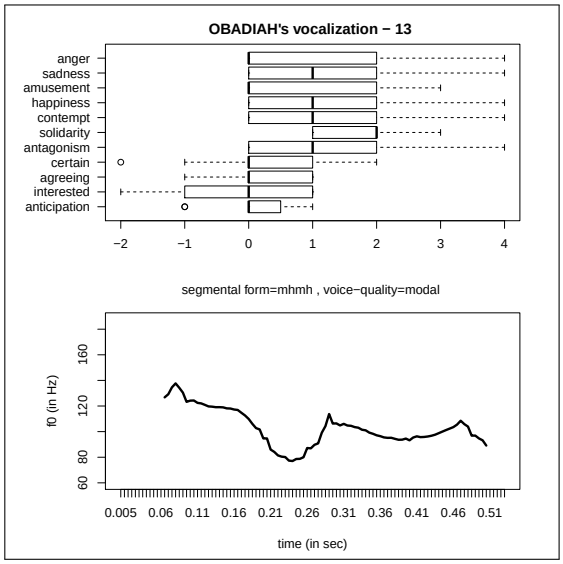

PRUDENCE's vocalization − 18
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
segmental form=tsgosh , voice−quality=modal
f0 (in Hz)
120
160
200
240
0.005
0.07
0.13
0.19
0.25
0.31
0.37
0.43
0.49
0.55
0.61
time (in sec)

Obadiah's vocalizations




OBADIAH's vocalization − 5
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
60 80 100 140
0.005 0.06 0.1 0.14 0.19 0.24 0.29 0.34 0.39 0.44
time (in sec)
OBADIAH's vocalization − 6
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=alright , voice−quality=modal
f0 (in Hz)
60 80 100 140
0.005 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46 0.51
time (in sec)
OBADIAH's vocalization − 7
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=mhmh_yeah , voice−quality=breathy
f0 (in Hz)
60 80 120 160
0.005 0.1 0.18 0.27 0.36 0.45 0.54 0.63 0.72 0.8 0.88
time (in sec)
OBADIAH's vocalization − 8
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=(sigh) , voice−quality=modal
f0 (in Hz)
60 80 120 160
0.005 0.08 0.15 0.22 0.29 0.36 0.43 0.5 0.56 0.63
time (in sec)


OBADIAH's vocalization − 9
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=ma , voice−quality=modal
f0 (in Hz)
60 80 100 140
0.005 0.05 0.09 0.13 0.17 0.21 0.25 0.29 0.33 0.37
time (in sec)
OBADIAH's vocalization − 10
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=yeah , voice−quality=modal
f0 (in Hz)
60 80 100 140
0.005 0.07 0.13 0.19 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
time (in sec)
OBADIAH's vocalization − 11
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=right , voice−quality=modal
f0 (in Hz)
60 80 100 140
0.005 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46 0.51
time (in sec)
OBADIAH's vocalization − 12
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2 −1 0 1 2 3 4
segmental form=mna_(snort) , voice−quality=modal
f0 (in Hz)
60 80 120 160
0.005 0.2 0.36 0.54 0.7 0.86 1 1.13 1.3 1.46 1.64 1.8
time (in sec)


OBADIAH's vocalization − 13
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
segmental form=mhmh , voice−quality=modal
f0 (in Hz)
60
80
120
160
0.005
0.06
0.11
0.16
0.21
0.26
0.31
0.36
0.41
0.46
0.51
time (in sec)

Appendix C

Vocalizations' meaning appropriateness

Based on the ratings received from several subjects using the web-based perception test, we obtained appropriateness of listener vocalizations for each of the meanings. This appendix presents the meaning appropriateness for each of the stimuli used in the multidimensional meaning annotation (described in Chapter 8). The tables in the following appendix shows meaning appropriateness for each of the SAL characters. The appropriateness of meaning-vocalization combination is represented using the following symbols:
$\circ$: vocalization is not appropriate for the meaning;
$\uparrow$ or $\downarrow$: vocalization is somewhat appropriate;
$\Uparrow$ or $\Downarrow$: vocalization is very appropriate for the meaning;
$\cdot$: the annotation has low agreement (we can not conclude on appropriateness);
$\downarrow$ and $\Downarrow$: negative sides of bipolar scales

Poppy's fixed segmental form set												
segmental form	intonation contour	voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	⌒	modal	○	○	·	·	○	⇑	○	⇑	⇑	⇑	·
yeah	‾	breathy	○	·	○	○	·	·	○	·	○	·	○
yeah	‾	modal	·	·	○	○	·	↑	·	·	○	·	○
yeah	‾	breathy	○	·	○	○	·	·	○	·	↑	○	○
yeah	~	modal	○	○	⇑	⇑	○	·	○	⇑	⇑	⇑	·
yeah	‾	modal	○	·	○	○	·	·	·	·	○	·	○
yeah	~	breathy	○	·	○	○	·	·	·	·	·	·	○
yeah	‾	modal	○	○	○	·	·	·	·	↑	○	·	○
yeah	‾	modal	○	○	○	○	·	·	·	·	↑	·	○
yeah	‾	tense	○	○	·	·	·	·	·	·	·	○	○
yeah	‾	modal	○	·	○	○	○	·	·	○	↑	↓	○
yeah	‾	modal	○	·	○	○	○	·	○	·	·	○	○
yeah	‾	modal	○	·	○	○	·	·	·	·	↓	·	○
yeah	⌒	modal	○	○	·	·	·	·	○	↑	⇑	↑	↑
yeah	‾	modal	○	○	·	·	·	·	·	⇑	·	↑	↑

Table C.1: Segmental form, intonation contour, voice quality and meaning appropriateness of Poppy's stimuli.

Poppy's fixed intonation set											
segmental form intonation contour voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
definitely ⁀ modal	○	·	·	·	·	⇑	·	·	·	○	↑
yeah ⁀ breathy	○	·	·	○	○	↑	·	·	○	·	○
gosh ⁀ modal	○	·	·	·	○	·	○	○	○	↑	·
(sigh) ⁀ breathy	○	·	○	○	·	·	·	·	·	·	·
yes ⁀ modal	·	○	·	·	·	·	·	⇑	·	·	·
mh ⁀ breathy	·	·	○	○	·	·	·	·	○	↓	·
really ⁀ tense	·	○	·	○	·	·	·	·	○	·	·

Table C.2: Segmental form, intonation contour, voice quality and meaning appropriateness of Poppy's stimuli.

Spike's fixed segmental form set												
segmental form	intonation contour	voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	—	creaky	○	·	○	○	↑	○	·	○	○	↓	·
yeah	—	modal	·	·	○	·	·	·	·	↓	·	·	·
yeah	—	modal	·	○	○	○	·	·	·	⇑	↑	↑	·
yeah	⁀	modal	·	○	○	○	·	↑	·	↑	↑	○	○
yeah	⁀	modal	·	○	○	○	·	⇑	○	⇑	·	↑	·
yeah	⁀	modal	○	○	○	○	·	·	○	○	↑	·	·
yeah	—	modal	·	○	○	○	·	·	·	⇑	·	○	○
yeah	⌡	creaky	○	·	○	○	·	·	·	·	·	·	·
yeah	⁀	modal	○	○	○	○	○	·	·	○	○	↑	↑
yeah	—	creaky	○	○	○	○	○	·	·	↑	↑	·	↑

Table C.3: Segmental form, intonation contour, voice quality and meaning appropriateness of Spikes's stimuli.

Spike's fixed intonation set											
segmental form intonation contour voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
aha ‿ modal	○	·	○	·	·	○	·	↑	○	↑	○
yeah_absolutely — creaky	·	·	·	·	·	·	○	⇑	⇑	↑	↑
oh_god_(gasp) ⁀ creaky	○	·	·	·	⇑	·	⇑	↑	·	○	○
hm ‿ modal	·	·	·	○	·	·	↑	·	↓	·	·
(snort) — breathy	·	·	·	·	·	·	·	↑	○	○	↑
mhm ‿ modal	·	·	○	○	·	↑	·	○	○	○	·
yeah_(gasp) ‿ modal	·	·	·	·	·	·	·	·	○	·	·
ah_I_see — modal	·	·	○	○	·	↑	·	·	·	○	·
I_see — modal	○	○	○	○	○	·	·	·	○	·	·

Table C.4: Segmental form, intonation contour, voice quality and meaning appropriateness of Spike's stimuli.

Prudence's fixed segmental form set												
segmental form	intonation contour	voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	⁀	modal	○	·	○	○	·	↑	○	↑	↑	○	○
yeah	⁀	modal	○	⇑	○	○	○	·	○	○	○	·	○
yeah	⁀	creaky	○	·	○	○	·	·	○	↑	↑	○	↑
yeah	⁀	modal	○	○	·	·	○	⇑	○	↑	↑	↑	·
yeah	⁀	modal	○	·	○	○	○	·	·	○	↑	·	·
yeah	⁀	modal	○	○	⇑	⇑	○	⇑	○	↑	↑	↑	·
yeah	⁀	creaky	○	·	○	○	○	·	○	○	○	↓	○
yeah	⁀	modal	○	·	○	○	·	↑	·	·	·	↓	○

Table C.5: Segmental form, intonation contour, voice quality and meaning appropriateness of Prudence's stimuli.

Prudence's fixed intonation set											
segmental form intonation contour voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
tsyes ‿ modal	·	·	·	·	·	·	·	↑	○	·	·
tsyeah ‿ modal	·	·	·	○	·	·	·	·	·	·	○
mhm ‿ modal	·	·	○	○	·	↑	·	·	○	·	·
yeah ‿ modal	·	·	○	○	·	·	·	↑	·	·	·
yes ‿ modal	·	·	○	○	·	·	·	○	○	○	↑
right ‿ modal	·	○	○	○	·	·	·	·	·	·	·
tsright ‿ modal	·	·	○	○	·	·	·	↑	·	·	○
aha ‿ modal	○	○	·	·	·	⇑	○	↑	↑	↑	↑
tsgosh ‿ modal	○	○	·	·	·	·	○	○	○	○	○

Table C.6: Segmental form, intonation contour, voice quality and meaning appropriateness of Prudence's stimuli.

Obadiah's fixed segmental form set												
segmental form	intonation contour	voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	—	modal	○	○	○	○	○	·	·	↑	↑	↓	○
yeah	—	breathy	○	⇑	○	○	○	·	·	○	○	·	○
yeah	—	modal	○	·	○	○	·	·	·	↓	·	↓	○
yeah	—	modal	·	○	○	○	○	·	·	↑	·	○	○
yeah	—	modal	○	·	○	○	○	·	·	↑	↑	·	○

Table C.7: Segmental form, intonation contour, voice quality and meaning appropriateness of Obadiah's stimuli.

Obadiah's fixed intonation set											
segmental form intonation contour voice quality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
alright — modal	·	·	·	·	·	·	·	↓	·	↓	·
mhmh_yeah — breathy	○	·	·	·	·	↑	↑	○	○	↓	○
(sigh) — modal	·	·	○	·	○	·	↑	○	○	↓	·
ma — modal	○	·	○	○	·	·	·	·	○	↓	↓
yeah — modal	○	·	○	○	·	·	·	○	·	↓	○
right — modal	·	·	·	·	·	·	·	○	↑	·	·
mna_(snort) — modal	·	·	·	·	⇑	·	↑	↓	·	·	○
mhmh — modal	·	○	·	·	·	↑	·	○	○	·	·

Table C.8: Segmental form, intonation contour, voice quality and meaning appropriateness of Obadiah's stimuli.

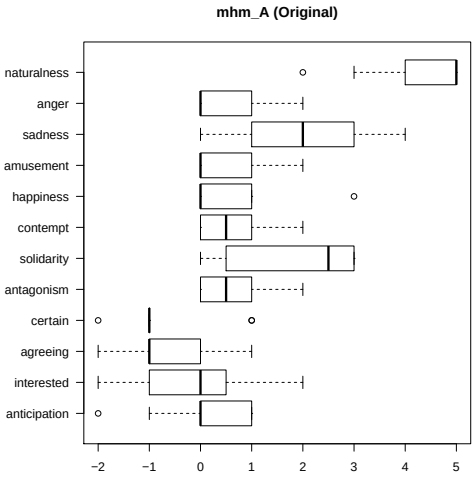
Appendix D

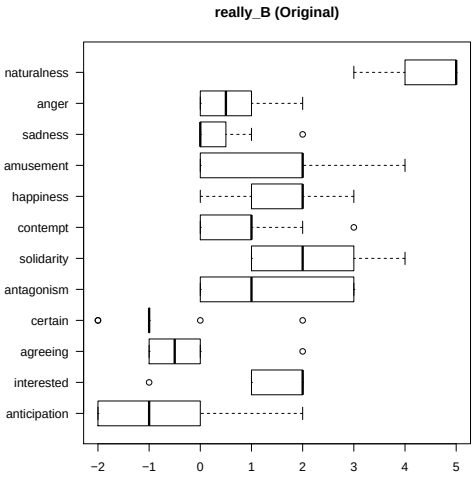
Ratings of evaluation test on signal modification techniques

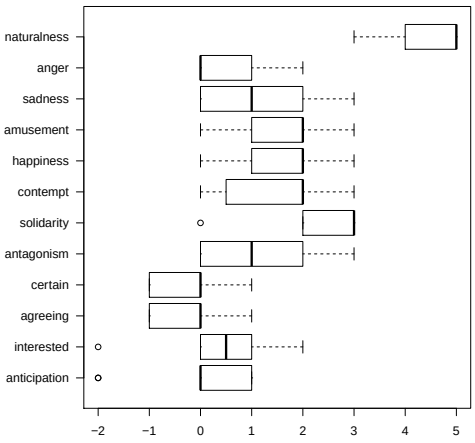
This appendix presents the ratings of evaluation test on the signal modification techniques such as *FDPSOLA, MLSA vocoding*, and *HNM vocoding*. The cross-combination of segmental forms and intonation contours of a selected set of stimuli is presented to several participants using a web-based perception test. The test consists of three original listener vocalizations and their cross-combinations of segmental forms and intonation contours with the help of signal modification techniques.

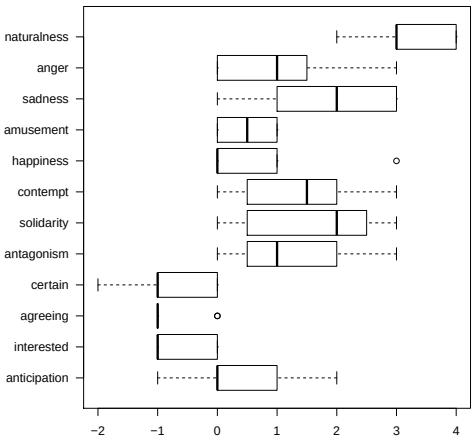
The original vocalizations are presented in Chapter 10. They are:

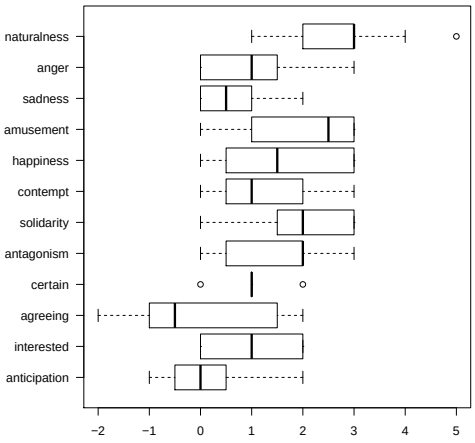
- *mhm_A* : segmental form *mhm* and its intonation contour *A*
- *really_B* : segmental form *really* and its intonation contour *B*
- *oh_C* : segmental form *oh* and its intonation contour *C*

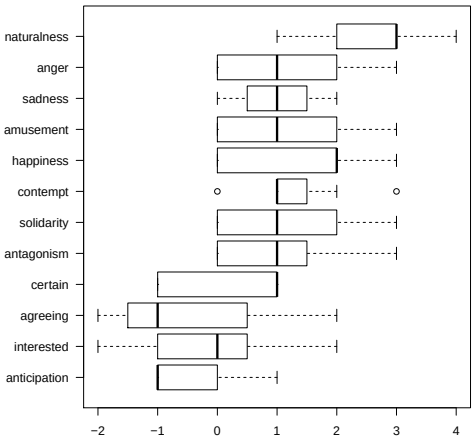

mhm_A (Original)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

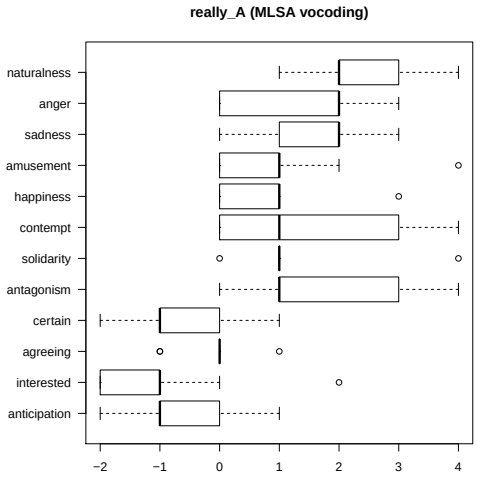

really_B (Original)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

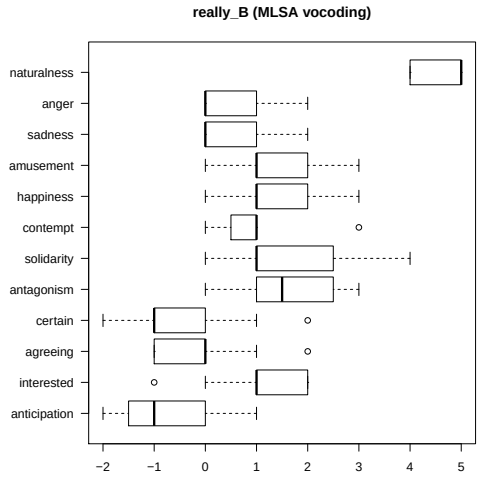

oh_C (Original)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

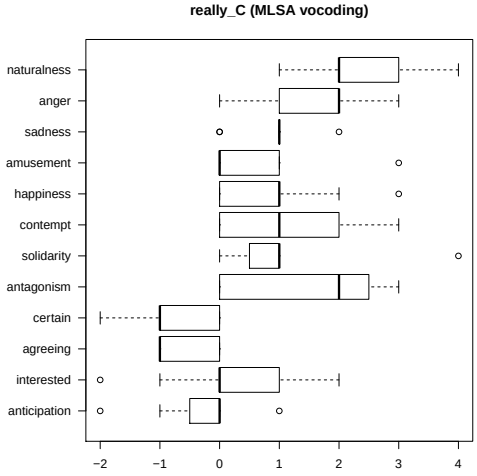

mhm_A (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

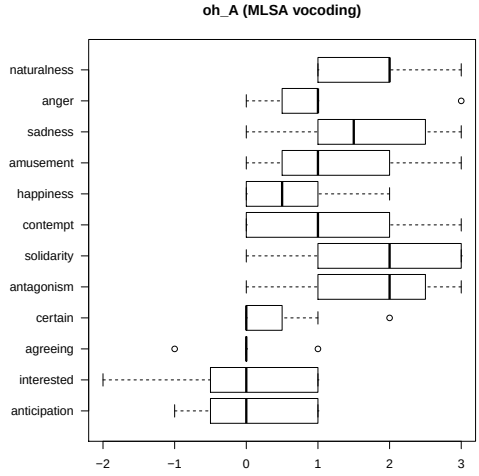

mhm_B (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

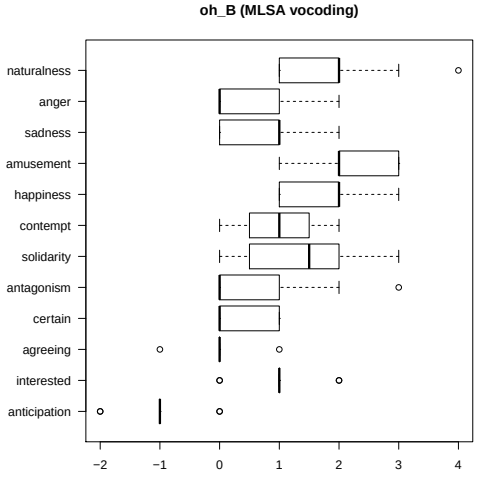

mhm_C (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

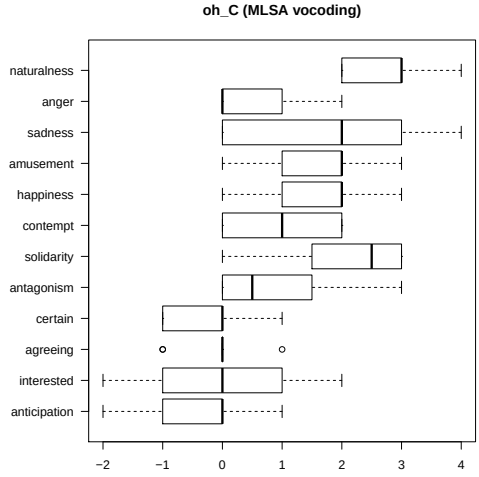

really_A (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

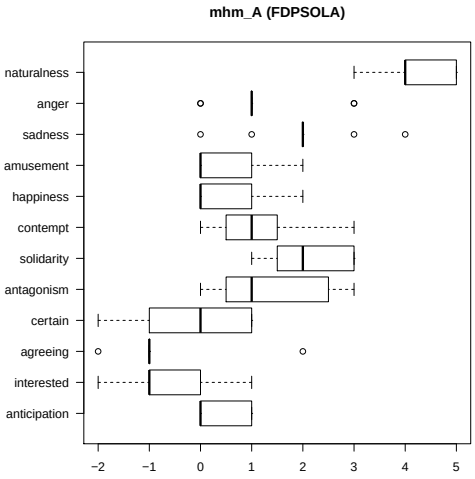

really_B (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

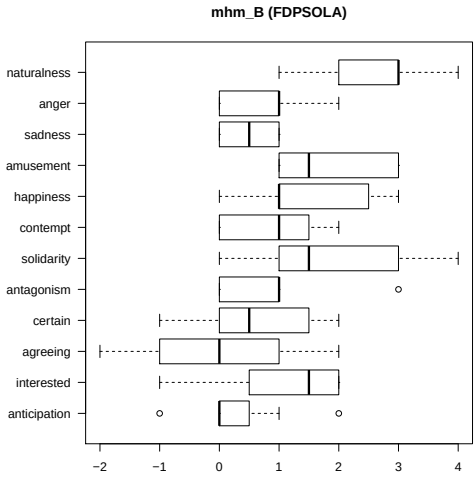

really_C (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

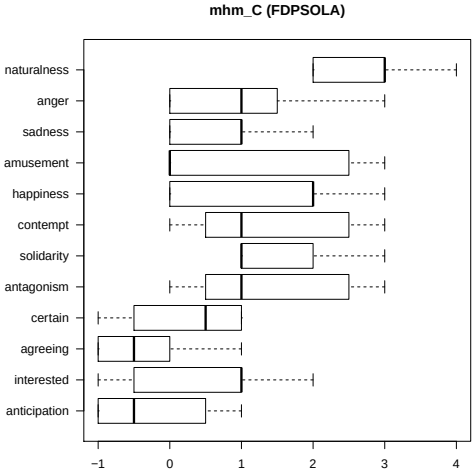

oh_A (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3

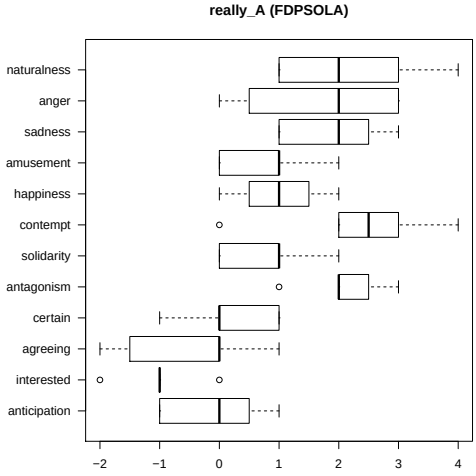

oh_B (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

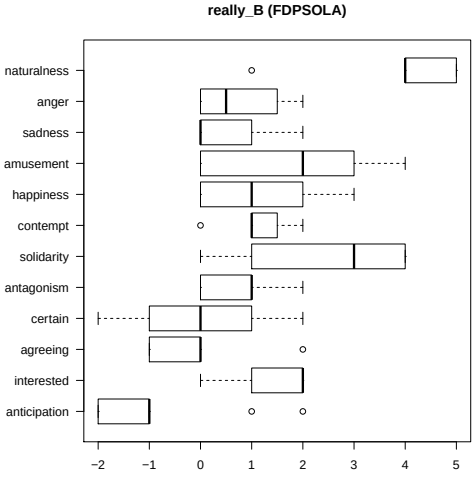

oh_C (MLSA vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

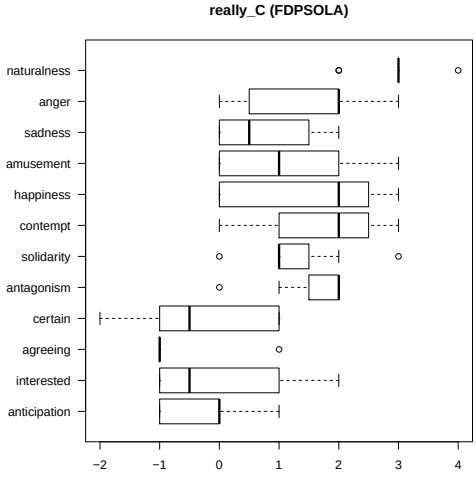

mhm_A (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

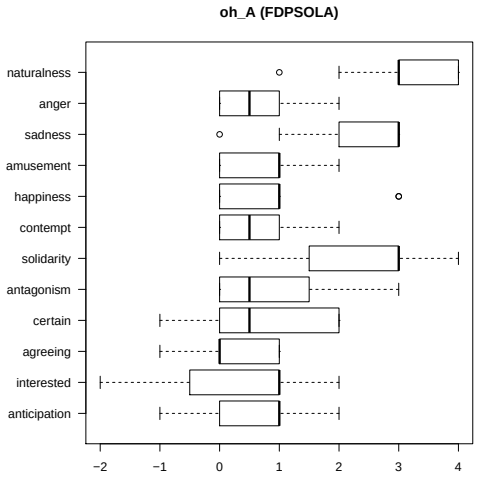

mhm_B (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

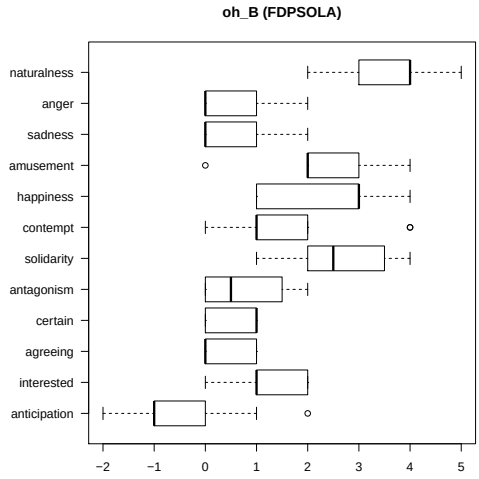

mhm_C (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−1
0
1
2
3
4

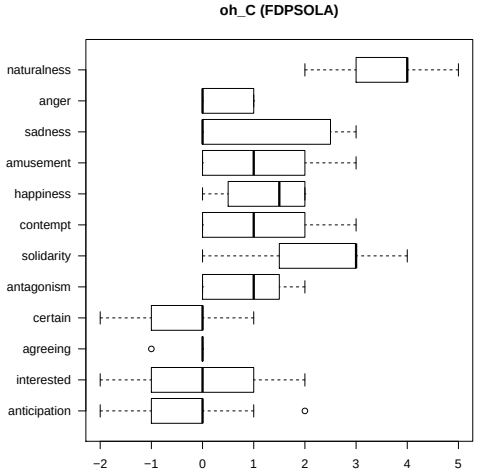

really_A (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

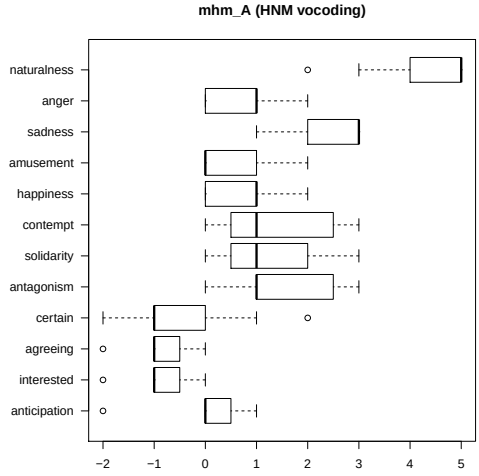

really_B (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

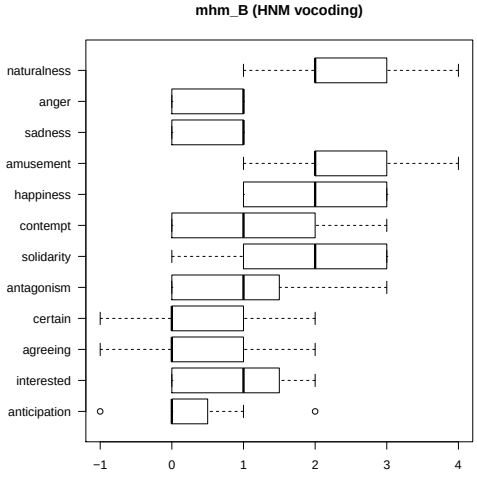

really_C (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

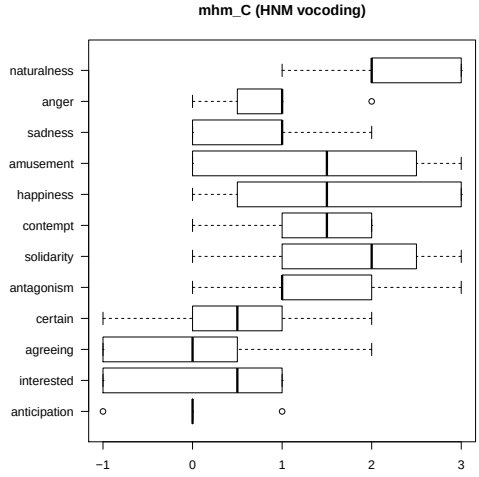

oh_A (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4

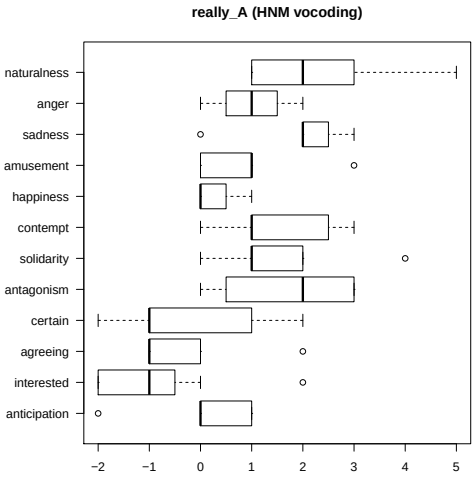

oh_B (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

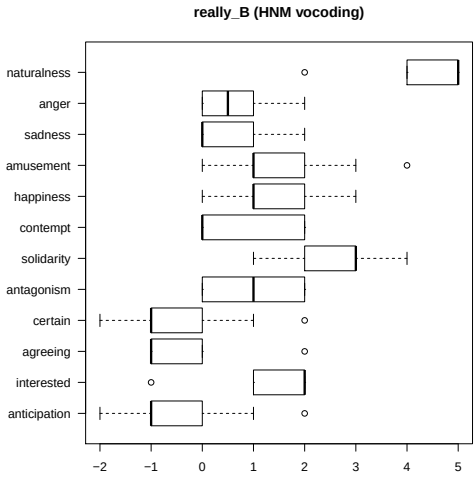

oh_C (FDPSOLA)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

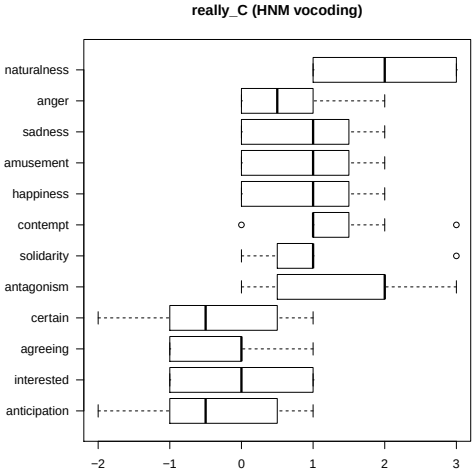

mhm_A (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

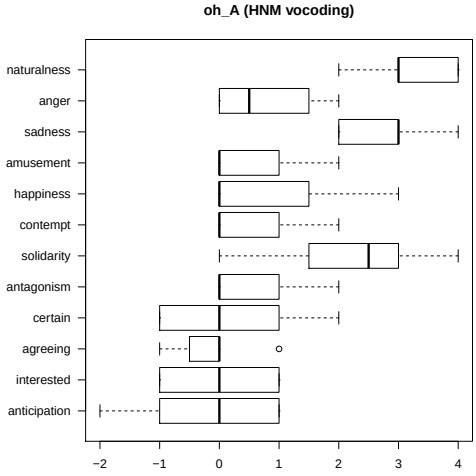

mhm_B (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−1
0
1
2
3
4

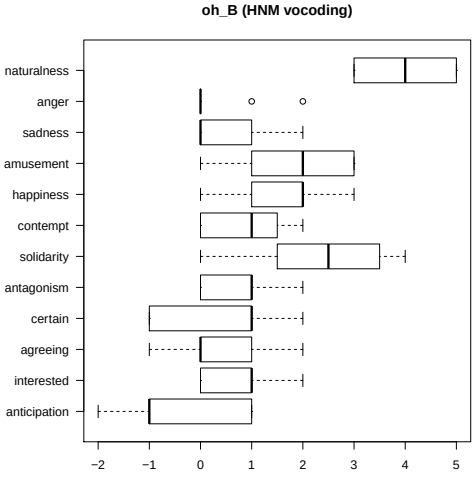

mhm_C (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−1
0
1
2
3

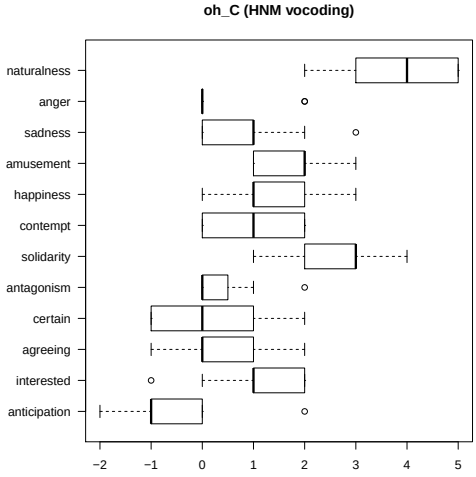

really_A (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5


really_B (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5


really_C (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3


oh_A (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4


oh_B (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5


oh_C (HNM vocoding)
naturalness
anger
sadness
amusement
happiness
contempt
solidarity
antagonism
certain
agreeing
interested
anticipation
−2
−1
0
1
2
3
4
5

Bibliography

Adell, J. (2009). “Analysis and modelling of conversational elements for speech synthesis”. PhD thesis. Barcelona, Spain: Universitat Politecnica de Catalunya (UPC).

Allwood, Jens, Joakim Nivre, and Elisabeth Ahlsén (1992). “On the Semantics and Pragmatics of Linguistic Feedback”. In: *Journal of Semantics* 9.1, pp. 1–26. DOI: 10.1093/jos/9.1.1.

Anderson, A. and T. Lynch (1988). *Listening*. Oxford: Oxford University Press.

Anumanchipalli, GK., K. Prahallad, and AW. Black (2011). “Festvox: Tools for Creation and Analyses of Large Speech Corpora”. In: *Proc. Workshop on Very Large Scale Phonetics Research*. UPenn, Philadelphia.

Atkinson, J.M. (1992). “Displaying neutrality: formal aspects of informal court proceedings”. In: *Talk at work: Interaction in institutional settings*, pp. 199–211.

Baillie, J.C. (2005). “Urbi: Towards a universal robotic low-level programming language”. In: *Proc. International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 820–825.

Bales, R.F. (1950). “Interaction process analysis: A method for the study of small groups”. In: *Cambridge Mass*.

Banavar, G. et al. (1999). “A case for message oriented middleware”. In: *Distributed Computing*, pp. 846–846.

Baron-Cohen, S. (1988). “Social and pragmatic deficits in autism: cognitive or affective?” In: *Journal of autism and developmental disorders* 18.3, pp. 379–402.

Baron-Cohen, S. et al. (2001). “The Reading the Mind in the Eyes Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism”. In: *Journal of Child Psychology and Psychiatry* 42.02, pp. 241–251.

Baron-Cohen, Simon et al. (2004). *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers.

Bavelas, J.B., L. Coates, and T. Johnson (2000). “Listeners as co-narrators.” In: *Journal of Personality and Social Psychology* 79.6, p. 941.

— (2002). “Listener responses as a collaborative process: The role of gaze”. In: *Journal of Communication* 52.3, pp. 566–580.

Bevacqua, E. (2009). “Computational model of listener behavior for Embodied Conversational Agents”. PhD thesis. University Paris 8.

Bevacqua, E. et al. (2007). “Facial Feedback Signals for ECAs”. In: *AISB 2007 Annual convention, workshop “Mindful Environments”*. Newcastle, UK, pp. 147–153.

Bevacqua, E. et al. (2010). “Multimodal Backchannels for Embodied Conversational Agents”. In: *Proc. Intelligent Virtual Agents*. Philadelphia, USA: Springer, pp. 194–200. DOI: 10.1007/978-3-642-15892-6_21.

Bilous, F.R. and R.M. Krauss (1988). “Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads”. In: *Language and Communication* 8.3-4, pp. 183 –194. DOI: 10.1016/0271-5309(88)90016-X.

Black, A. and P. Taylor (1997). “Automatically clustering similar units for unit selection in speech synthesis”. In: *Proceedings of EUROSPEECH*. Vol. 2, pp. 601–604.

Black, A., P. Taylor, and R. Caley (1998). *The Festival speech synthesis system*. url: http://www.cstr.ed.ac.uk/projects/festival.

Black, A.W. and K.A. Lenzo (2003). *Building synthetic voices*. Language Technologies Institute, Carnegie Mellon University. url: `http://festvox.org/bsv`.

Black, A.W., H. Zen, and K. Tokuda (2007). "Statistical parametric speech synthesis". In: *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 4. IEEE, pp. IV–1229.

Boersma, P. and D. Weenink (2010). *Praat: doing phonetics by computer*. Computer program, version 5, url: `www.praat.prg`.

Bonafonte, A. et al. (2006). "Ogmios: The UPC text-to-speech synthesis system for spoken translation". In: *Proc. TC-STAR Workshop on Speech-to-Speech Translation*.

Bonafonte, A. et al. (2008). "The UPC TTS system description for the 2008 blizzard challenge". In: *Proc of the Blizzard Challenge, Brisbane, Australia, September*.

Bratman, M.E. (1987). *Intention, plans, and practical reason*. Center for the Study of Language and Information - The David Hume Series.

— (1990). *What is intention?* Ed. by P. R. Cohen, J. L. Morgan, and M. E. Pollack. The MIT Press: Cambridge, MA, pp. 15–31.

Brown, P. and S.C. Levinson (1987). *Politeness: Some universals in language usage*. Vol. 4. Cambridge University Press.

Brunner, L.J. (1979). "Smiles can be back channels." In: *Journal of Personality and Social Psychology* 37.5, p. 728.

Bublitz, W. (1988). *Supportive fellow-speakers and cooperative conversations*. Cambridge University Press.

Bühler, K. (1934). *Sprachtheorie*. Stuttgart, Germany: Gustav Fischer Verlag.

Campbell, N. (2005). "Developments in corpus-based speech synthesis: Approaching natural conversational speech". In: *IEICE transactions on information and systems* 88.3, pp. 376–383.

— (2006). "Conversational speech synthesis and the need for some laughter". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 14.4, pp. 1171–1178.

Carletta, J. et al. (2006). "The AMI meeting corpus: A pre-announcement". In: *Machine Learning for Multimodal Interaction*, pp. 28–39.

Carlson, R., K. Gustafson, and E. Strangert (2006). "Modelling hesitation for synthesis of spontaneous speech". In: *Proc. Speech Prosody*.

Cassell, J. et al. (1999). "Embodiment in conversational interfaces: Rea". In: *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*. ACM, pp. 520–527.

Clancy, P.M. et al. (1996). "The conversational use of reactive tokens in English, Japanese, and Mandarin* 1". In: *Journal of Pragmatics* 26.3, pp. 355–387.

Clark, H.H. and M.A. Krych (2004). "Speaking while monitoring addressees for understanding* 1". In: *Journal of Memory and Language* 50.1, pp. 62–81.

Cowie, Roddy, Naomi Sussman, and Aaron Ben-Ze'ev (2011). "Emotion: Concepts and Definitions". In: *Emotion-Oriented Systems*, pp. 9–30.

Dittmann, A.T. and L.G. Llewellyn (1968). "Relationship between vocalizations and head nods as listener responses." In: *Journal of personality and social psychology* 9.1, p. 79.

Douglas-Cowie, E., R. Cowie, and M. Schröder (2003). "The description of naturally occurring emotional speech". In: *Proc. 15th Internat. Conf. on Phonetic Sciences, Barcelona, Spain*. Citeseer, pp. 2877–2880.

Douglas-Cowie, Ellen et al. (2008). "The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation". In: *LREC Workshop on Corpora for Research on Emotion and Affect*. Marrakech, Morocco, pp. 1–4.

Drummond, K. and R. Hopper (1993). "Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency." In: *Research on Language and Social Interaction* 26.2, pp. 157–77.

Duncan, S. (1974). "On the structure of speaker–auditor interaction during speaking turns". In: *Language in society* 3.02, pp. 161–180.

Dutoit, T. et al. (1996). "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes". In: *Proc. Fourth International Conference on Spoken Language Processing*. Vol. 3. IEEE, pp. 1393–1396.

Ekman, P. (1999). *Basic emotions*. Ed. by T. Dalgleish and T. Power. Sussex, U.K.: John Wiley & Sons, Ltd., pp. 45–60.

Feke, M.S. (2003). "Effects of Native-Language and Sex on Back-Channel Behavior". In: *Selected Proceedings from the First Workshop on Spanish Sociolinguistics. Somerville, MA: Cascadilla Proceedings Project*, pp. 96–106.

Fernandez, R. and B. Ramabhadran (2007). "Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis". In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pp. 34–39.

Fishman, P. (1978). "Interaction: The work women do". In: *Social problems* 24, pp. 397–406.

French, P. and J. Local (1983). "Turn-competitive incomings". In: *Journal of Pragmatics* 7.1, pp. 17–38.

Fries, C.C. (1952). *The Structure of English (New York)*.

Fujii, Kei, Hideki Kashioka, and Nick Campbell (2003). "Target cost of F_0 based on polynomial regression in concatenative speech synthesis". In: *Proceedings of the 15th International Conference of Phonetic Sciences*. Barcelona, Spain, pp. 2577–2580.

Fujimoto, D.T. (2009). "Listener responses in interaction: A case for abandoning the term, backchannel".

Garcés Conejos, P. and P. Bou Franch (2004). "A pragmatic account of listenership: implications for foreign/second language teaching". In: *Revista alicantina de estudios ingleses* 17, pp. 81–102.

Gardner, Rod (2002). *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins Publishing Co.

Gebhard, P. et al. (2008). "IDEAS4Games: building expressive virtual characters for computer games". In: *Proc. Intelligent Virtual Agents*. Springer, pp. 426–440.

Goodwin, C. (1986). "Between and within: Alternative sequential treatments of continuers and assessments". In: *Human Studies* 9.2, pp. 205–217.

Gratch, J. et al. (2007). "Creating rapport with virtual agents". In: *Proc. Intelligent Virtual Agents*. Springer, pp. 125–138.

Härdle, W. and L. Simar (2007). *Applied multivariate statistical analysis*. Springer.

Haugh, M. (2008). "Intention in pragmatics". In: *Intercultural Pragmatics* 5.2, pp. 99–110.

Heinz, B. (2003). "Backchannel responses as strategic responses in bilingual speakers' conversations". In: *Journal of pragmatics* 35.7, pp. 1113–1142.

Heylen, D., A. Nijholt, and M. Poel (2007). "Generating nonverbal signals for a sensitive artificial listener". In: *Verbal and Nonverbal Communication Behaviours*, pp. 264–274.

Heylen, D. et al. (2007). "Searching for prototypical facial feedback signals". In: *Proc. Intelligent Virtual Agents*. Springer, pp. 147–153.

Heylen, D. et al. (2008). "Why conversational agents do what they do? Functional representations for generating conversational agent behavior". In: *The First Functional Markup Language Workshop. Estoril, Portugal*.

Hirschman, L. (1994). "Female–male differences in conversational interaction". In: *Language in society* 23.03, pp. 427–442.

Holmes, J. (1997). "Story-telling in New Zealand women's and men's talk". In: *Gender and discourse* 131.

Hunecke, A. (2007). "Optimal design of a speech database for unit selection synthesis". Unpublished Diploma Thesis (Diplomarbeit), Universität des Saarlandes, Saarbrücken, Germany.

Hunt, A.J. and A.W. Black (1996). "Unit selection in a concatenative speech synthesis system using a large speech database". In: *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 373–376.

Iida, A. and N. Campbell (2003). "Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders". In: *International Journal of Speech Technology* 6.4, pp. 379–392.

Imai, S. (1983). "Cepstral analysis synthesis on the mel frequency scale". In: *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 8. IEEE, pp. 93–96.

Inanoglu, Z. and S. Young (2009). "Data-driven emotion conversion in spoken English". In: *Speech Communication* 51.3, pp. 268–283.

Jefferson, G. (1983). "Caveat Speaker: Preliminary Notes on Recipient Topic-Shift Implicature." In: *Research on Language and Social Interaction* 26.1, pp. 1–30.

Kasper, G. (1989). “Interactive procedures in interlanguage discourse”. In: *Contrastive pragmatics* 3, pp. 189–230.

Kendon, A. (1967). “Some functions of gaze-direction in social interaction.” In: *Acta psychologica* 26.1, pp. 22–63.

Kobayashi, T. et al. (2009). *Speech signal processing toolkit (SPTK), Version 3.3*. http://sp-tk.sourceforge.net/.

Kopp, S. et al. (2008). “Modeling embodied feedback with virtual humans”. In: *Modeling communication with robots and virtual humans*, pp. 18–37.

Krauss, R.M. and S. Weinheimer (1966). “Concurrent feedback, confirmation, and the encoding of referents in verbal communication.” In: *Journal of Personality and Social Psychology* 4.3, p. 343.

Kruijff-Korbayová, I. et al. (2011). “An Event-Based Conversational System for the Nao Robot”. In: *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*. Springer, pp. 125–132.

Langer, E.J. (1992). “Matters of mind: Mindfulness/mindlessness in perspective”. In: *Consciousness and Cognition* 1.3, pp. 289–305.

Laver, J. (1991). *The gift of speech: papers in the analysis of speech and voice*. Edinburgh University Press.

Leet-Pellegrini, H.M. (1980). “Conversational dominance as a function of gender and expertise”. In: *Language: Social psychological perspectives*, pp. 97–104.

Levinson, S.C. (1983). *Pragmatics*. Cambridge University Press.

Lyons, J. (1968). *Introduction to theoretical linguistics*. Cambridge Univ Pr.

Maatman, R., J. Gratch, and S. Marsella (2005). “Natural behavior of a listening agent”. In: *Proc. Intelligent Virtual Agents*. Springer, pp. 25–36.

Mancini, M. et al. (2008). “Greta: a SAIBA compliant ECA system”. In: *Proc. 3e Workshop sur les Agents Conversationnels Animées*.

Manusov, V. and A.R. Trees (2002). “Are You Kidding Me?: The Role of Nonverbal Cues in the Verbal Accounting Process”. In: *Journal of communication* 52.3, pp. 640–656.

Maynard, S.K. (1997). “Analyzing Interactional Management in Native/Non-Native English Conversation: A Case of Listener Response.” In: *IRAL* 35.1, pp. 37–60.

McCarthy, M. (2003). “Talking back: “Small" interactional response tokens in everyday conversation”. In: *Research on Language and Social Interaction* 36, pp. 33–63.

McCree, A. V. and T. P. Barnwell (1995). “A mixed excitation LPC vocoder model for low bit rate speech coding”. In: *IEEE Transactions on Speech and Audio Processing* 3.4, pp. 242–249.

McKeown, Gary et al. (2010). “The SEMAINE corpus of emotionally coloured character interactions”. In: *Proc. IEEE International Conference on Multimedia & Expo*. Singapore, pp. 1079–1084.

Miyazaki, N. et al. (1998). *A study on pitch pattern generation using HMMs based on multi-space probability distributions*. Tech. rep. IEICE.

Morency, L.P., I. de Kok, and J. Gratch (2010). “A probabilistic multimodal approach for predicting listener backchannels”. In: *Autonomous Agents and Multi-Agent Systems* 20.1, pp. 70–84.

Moulines, E. and W. Verhelst (1995). “Time-domain and frequency-domain techniques for prosodic modification of speech”. In: *Speech coding and synthesis*. Ed. by W. Kleijn and K. Paliwal. Elsevier, pp. 519–555.

Nass, C. and Y. Moon (2000). "Machines and mindlessness: Social responses to computers". In: *Journal of Social Issues* 56.1, pp. 81–103.

Niewiadomski, R. et al. (2009). "Greta: an interactive expressive ECA system". In: *Proc. International Conference on Autonomous Agents and Multiagent Systems*. Vol. 2, pp. 1399–1400.

Nijholt, A. et al. (2008). "Mutually coordinated anticipatory multimodal interaction". In: *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pp. 70–89.

Odell, J.J. (1995). "The use of context in large vocabulary speech recognition". PhD thesis. Queens' College.

Ostermann, J. (2002). "Face Animation in MPEG-4". In: *MPEG-4 facial animation: the standard, implementation and applications*, pp. 17–55.

Pammi, S., M. Charfuelan, and M. Schröder (2009). "Quality control of automatic labelling using HMM-based synthesis". In: *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

— (2010). "Multilingual voice creation toolkit for the MARY TTS platform". In: *Proc. Int. Conf. Language Resources and Evaluation, Malta*.

Pammi, S. and M. Schröder (2009). "A corpus based analysis of backchannel vocalizations". In: *Proc. Interdisciplinary Workshop on Laughter and other Interactional Vocalisations in Speech*. Berlin, Germany.

Pammi, S. et al. (2010). "Synthesis of listener vocalisations with imposed intonation contours". In: *Proc. Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*. Kyoto, Japan.

Pelachaud, C. (2005). "Multimodal expressive embodied conversational agents". In: *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, pp. 683–689.

Pfleger, Norbert and Jan Alexandersson (2004). "Modeling non-verbal behavior in multimodal conversational systems". In: *Information Technology* 46.6, pp. 341–345.

Poggi, I. et al. (2005). "Greta. a believable embodied conversational agent". In: *Multimodal intelligent information presentation*, pp. 3–25.

Reidsma, D. et al. (2011). "Continuous Interaction with a Virtual Human". In: *Journal on Multimodal User Interfaces*.

Richards, J.C. (1983). "Communicative needs in foreign language learning1". In: *ELT Journal* 37.2, p. 111.

Schegloff, E. (1982a). "Discourse as an interactional achievement". In: *Analyzing discourse: Text and talk*, pp. 71–93.

Schegloff, E.A. (1982b). "Discourse as an interactional achievement: Some uses of uh huhand other things that come between sentences". In: *Analyzing discourse: Text and talk* 71.93, pp. 361–82.

Scherer, K. R. (1988). "On the symbolic functions of vocal affect expression". In: *Journal of Language and Social Psychology*, 7:79–100.

Scherer, Klaus R. (2005). "What are emotions? And how can they be measured?" In: *Social Science Information* 44.4, pp. 695–729.

Schiffrin, D. (1988). *Discourse markers*. Vol. 5. Cambridge University Press.

Schröder, M. (2006). "Expressing degree of activation in synthetic speech". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 14.4, pp. 1128–1136.

Schröder, M. (2007). "Interpolating expressions in unit selection". In: *Affective Computing and Intelligent Interaction*, pp. 718–720.

— (2009). "Expressive speech synthesis: Past, present, and possible futures". In: *Affective information processing*, pp. 111–126.

— (2010). "The SEMAINE API: towards a standards-based framework for building emotion-oriented systems". In: *Advances in Human-Computer Interaction* 2010.

Schröder, M., D. K. J. Heylen, and I. Poggi (2006). "Perception of non-verbal emotional listener feedback". In: *Proc. Speech Prosody 2006, Dresden*. Ed. by R. Hoffmann and H. Mixdorff. Vol. 40. Studientexte zur Sprachkommunication. Dresden: TUDpress Delft, pp. 43–46. ISBN: 3-938863-57-9.

Schröder, M. and A. Hunecke (2007). "MARY TTS participation in the Blizzard Challenge 2007". In: *Proceedings of the Blizzard Challenge 2007*.

Schröder, M., A. Hunecke, and S. Krstulovic (2006). "OpenMary–open source unit selection as the basis for research on expressive synthesis". In: *Proc. Blizzard Challenge*. Vol. 6.

Schröder, M. and J. Trouvain (2003). "The German text-to-speech synthesis system MARY: A tool for research, development and teaching". In: *International Journal of Speech Technology* 6.4, pp. 365–377.

Schröder, M. et al. (2008a). "Enhancing Animated Agents in an Instrumented Poker Game". In: *KI 2008: Advances in Artificial Intelligence*.

Schröder, M. et al. (2008b). "Towards responsive sensitive artificial listeners". In: *Proceedings of the 4th International Workshop on Human-Computer Conversation*.

Schröder, M. et al. (2009). "A demonstration of audiovisual sensitive artificial listeners". In: *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 1–2.

Schweitzer, A. et al. (2003). "Restricted unlimited domain synthesis". In: *Proceedings of Eurospeech*. Citeseer, pp. 1321–1324.

Schweitzer, A. et al. (2006). "Multimodal speech synthesis". In: *SmartKom: Foundations of Multimodal Dialogue Systems*, pp. 411–435.

Sevin, E. de et al. (2010). "A Multimodal Listener Behaviour Driven by Audio Input". In: *Proc. International Workshop on Interacting with ECAs as Virtual Characters*, p. 34.

Silverman, Kim et al. (1992). "ToBI: A standard for labeling English prosody". In: *Proceedings of the 2nd International Conference of Spoken Language Processing*. Banff, Canada, pp. 867–870.

Sinclair, J.M.H. and M. Coulthard (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press London.

Sjölander, K. (2006). *The Snack Sound Toolkit*. http://www.speech.kth.se/snack (accessed on 25th June, 2011).

Sperber, D. and D. Wilson (1995). *Relevance: Communication and cognition*. Wiley-Blackwell.

Steiner, I. et al. (2010). "Symbolic vs. acoustics-based style control for expressive unit selection". In: *Proc. Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*.

Strangert, E. and R. Carlson (2006). "On modelling and synthesis of conversational speech". In: *Proc. Nordic Prosody IX, 2004*, pp. 255–264.

Stubbe, M. (1998). "Are you listening? Cultural influences on the use of supportive verbal feedback in conversation". In: *Journal of Pragmatics* 29.3, pp. 257–289.

Stylianou, Yiannis (1996). "Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modification". PhD Thesis. École nationale supérieure des télécommunication.

Taylor, P. et al. (1999). "Edinburgh speech tools library". In: *System Documentation Edition* 1, pp. 1994–1999.

Thórisson, K.R. (1996). "Communicative humanoids: a computational model of psychosocial dialogue skills". PhD thesis. Massachusetts Institute of Technology.

— (2002). "Natural turn-taking needs no manual: Computational theory and model, from perception to action". In: *Multimodality in language and speech systems*, pp. 173–207.

Thórisson, K.R. et al. (2005). "Whiteboards: Scheduling blackboards for semantic routing of messages & streams". In: *AAAI Workshop on Modular Construction of Human-Like Intelligence*, pp. 8–15.

Toda, T. and K. Tokuda (2005). "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis". In: *Proc. Ninth European Conference on Speech Communication and Technology*.

Tokuda, K., H. Zen, and A.W. Black (2002). "An HMM-based speech synthesis system applied to English". In: *Proc. IEEE Workshop on Speech Synthesis*. IEEE, pp. 227–230.

Tokuda, K. et al. (2000). "Speech parameter generation algorithms for HMM-based speech synthesis". In: *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1315–1318.

Tokuda, K. et al. (2008). *The HMM-based speech synthesis system (HTS) Version 2.1*. Online: http://hts.sp.nitech.ac.jp/ (accessed on 25th June, 2011).

Tokuda, K. et al. (2010). *HMM-based Speech Synthesis System (HTS)*. http://hts.sp.nitech.ac.jp/.

Tottie, G. (1991). "Conversational style in British and American English: The case of backchannels". In: *English corpus linguistics*, pp. 254–271.

Versloot, CA (2005). *What do Listeners do? A Simple Annotation Schema*. Tech. rep. Enschede: University of Twente (HMI).

Vilhjálmsson, H. et al. (2007). "The behavior markup language: Recent developments and challenges". In: *Proc. Intelligent Virtual Agents*. Springer, pp. 99–111.

Ward, N. and W. Tsukahara (2000). "Prosodic features which cue back-channel responses in English and Japanese". In: *Journal of Pragmatics* 32.8, pp. 1177–1207.

Ward, Nigel (2006). "Non-lexical conversational sounds in American English". In: *Pragmatics & Cognition* 14.1, pp. 131–184.

Xudong, D. (2009). "Listener response". In: *The pragmatics of interaction* 4, pp. 104–124.

Yamagishi, J. et al. (2003). "Modeling of various speaking styles and emotions for HMM-based speech synthesis". In: *Proc. Eighth European Conference on Speech Communication and Technology*.

Yamagishi, J. et al. (2007). "Model adaptation approach to speech synthesis with diverse voices and styles". In: *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 4. IEEE.

Yngve, V. H. (1970). "On getting a word in edgewise". In: *Chicago Linguistic Society. Papers from the 6th regional meeting*. Vol. 6, pp. 567–577.

Yoshimura, T. et al. (1998). "Duration modeling for HMM-based speech synthesis". In: *Proc. Fifth International Conference on Spoken Language Processing*.

— (1999). "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis". In: *Proceedings of Eurospeech 1999*. Budapest, Hungary.

Yoshimura, T. et al. (2001a). “Mixed excitation for HMM-based speech synthesis”. In: *Proc. Seventh European Conference on Speech Communication and Technology*.

— (2001b). “Mixed excitation for HMM-based speech synthesis”. In: *Proc. Eurospeech 2001*. Aalborg, Denmark.

Zen, H., K. Tokuda, and A.W. Black (2009). “Statistical parametric speech synthesis”. In: *Speech Communication* 51.11, pp. 1039–1064.

Zovato, E. et al. (2004). “Towards emotional speech synthesis: A rule based approach”. In: *Proc. 5th ISCA Speech Synthesis Workshop*, pp. 219–220.