Synthesis and Evaluation of Conversational Characteristics in Speech Synthesis

Sebastian Andersson

Doctor of Philosophy

University of Edinburgh

2013

Abstract

Conventional synthetic voices can synthesise neutral read aloud speech well. But, to make synthetic speech more suitable for a wider range of applications, the voices need to express more than just the word identity. We need to develop voices that can partake in a conversation and express, e.g. agreement, disagreement, hesitation, in a natural and believable manner.

In speech synthesis there are currently two dominating frameworks: unit selection and HMM-based speech synthesis. Both frameworks utilise recordings of human speech to build synthetic voices. Despite the fact that the content of the recordings determines the segmental and prosodic phenomena that can be synthesised, surprisingly little research has been made on utilising the corpus to extend the limited behaviour of conventional synthetic voices. In this thesis we will show how natural sounding conversational characteristics can be added to both unit selection and HMM-based synthetic voices, by adding speech from a spontaneous conversation to the voices.

We recorded a spontaneous conversation, and by manually transcribing and selecting utterances we obtained approximately two thousand utterances from it. These conversational utterances were rich in conversational speech phenomena, but they lacked the general coverage that allows unit selection and HMM-based synthesis techniques to synthesise high quality speech. Therefore we investigated a number of blending approaches in the synthetic voices, where the conversational utterances were augmented with conventional read aloud speech.

The synthetic voices that contained conversational speech were contrasted with conventional voices without conversational speech. The perceptual evaluations showed that the conversational voices were generally perceived by listeners as having a more conversational style than the conventional voices. This conversational style was largely due to the conversational voices' ability to synthesise utterances that contained conversational speech phenomena in a more natural manner than the conventional voices. Additionally, we conducted an experiment that showed that natural sounding conversational characteristics in synthetic speech can convey pragmatic information, in our case an impression of certainty or uncertainty, about a topic to a listener. The conclusion drawn is that the limited behaviour of conventional synthetic voices can be enriched by utilising conversational speech in both unit selection and HMM-based speech synthesis.

Acknowledgements

- Matthew Aylett: for giving me the essential support and supervision to complete this work.
- Robert Clark and Junichi Yamagishi: for supervision and collaboration.
- All colleagues at CSTR and other collaborators, in particular Kallirroi Georgila and David Traum.
- Oliver Watts and Avril Heron: for assisting me with those important things that are very difficult to do from abroad, such as printing and submitting the thesis.
- This work was made possible through the financial support from the Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sebastian Andersson*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of this thesis is to produce synthetic speech that can express conversational characteristics in a more natural and believable manner than conventional synthetic voices. Current speech synthesis techniques use a corpora of speech data to synthesise new utterances. Our approach is to augment the conventional database of neutrally read aloud data with speech from a spontaneous conversation, in order to achieve our goal of synthesising speech that exhibits conversational characteristics.

1.1 Utilising Richer Speech Resources

Unit selection and HMM-based synthetic voices can synthesise neutral read aloud speech well (see e.g. King and Karaiskos, 2009). For many applications, such as GPS systems or reading aloud text books, an intelligible read aloud speaking style is sufficient to provide a user with relevant information. But applications created to portray a believable character require synthetic voices that can express more than just propositional information. The characters need voices that can give an impression of being engaged in an interactive exchange by signalling turn-taking behaviour and provide backchannels, give an impression of self-motivation and intent by signalling agreement, disagreement, hesitation, et cetera (Loyall, 1997; Traum et al., 2008; Romportl et al., 2010). The challenge for speech synthesis in making synthetic voices suitable for believable characters is therefore not to make the synthetic voices capable of synthesising more natural-sounding propositional information, but to make a wider range of speech phenomena sound natural.

To build synthetic voices capable of expressing a wider range of speech phenomena than just propositional information we could attempt to generate the acoustic proper-

ties of these other speech phenomena with some signal processing method, or we can use the unit selection (Aylett and Pidcock, 2007) and HMM-based (Zen et al., 2007) speech synthesis techniques to learn the segmental and prosodic properties of a wider range of speech phenomena directly from speech data that contains them. Speech phenomena with global acoustic properties can be modelled using utterance level signal processing, e.g. the modelling of "happy" utterances in Romportl et al. (2010) by increasing the speaking rate and raising the F0 of neutral utterances. However, many other speech phenomena are of more local character, e.g. the phonetic properties of different phonemes in different syllable or utterance positions. Unit selection and HMM-based speech synthesis techniques have proven successful in learning these local properties directly from recordings of human speech and producing high quality synthetic utterances (Karaiskos et al., 2008; King and Karaiskos, 2009, 2010). The unit selection and HMM-based speech synthesis frameworks are formulated to preserve the segmental and prosodic properties of the recorded speech (Clark et al., 2007; Aylett and Pidcock, 2007; Zen et al., 2009). In order to build synthetic voices that are more suitable for interactive believable characters with these techniques one solution is to attempt to learn the segmental and prosodic properties from speech resources that contain a richer variety of the speech phenomena associated with human interaction than the more conventionally used speech resources of carefully and neutrally read aloud isolated sentences. This is the approach taken by this thesis.

Spontaneous conversations contain a rich variety of the speech phenomena of human everyday communication, including propositional information, but also discourse markers, filled pauses and backchannels (Clark, 1996). The structure and content of conversations will be described in more detail in sections 2.1 and 2.2. Discourse markers (e.g. *okay*, *you know*, *'cause*) and filled pauses (*um* and *uh*) are frequently used in conversation to signal the beginning, continuation or end of a conversational turn, as well as to signal affective content such as agreement or hesitation (Schiffrin, 1987; Jurafsky et al., 1998; Clark and Fox Tree, 2002). For example, signalling agreement by beginning an utterance with *yeah* or *oh yeah* (Jurafsky et al., 1998). However, utilising speech from a spontaneous conversation directly to build synthetic voices is difficult compared to the conventional approach of using neutrally read aloud sentences. It is difficult, firstly because conventional sub-word speech synthesis requires a segmental level match between audio and text which cannot be obtained automatically from conversations, and secondly because conversations contain an abundance of speech phenomena that are currently not modelled well in speech synthesis, e.g. heavily re-

duced, mispronounced or fragmented words, mumbling, interrupted utterances, sighs, coughs and laughter.

A more controllable alternative to speech from a conversation is to use acted speech, as in Gustafsson and Sjölander (2004); Cadic and Segalen (2008); Romportl et al. (2010); Adell et al. (2010). Although acted speech is useful for particular applications, the quality of a particular speech phenomenon, e.g. the hesitation and laughter in Cadic and Segalen (2008), will depend on how well the actor can act, whereas in speech from a conversation the acoustic properties of the speech phenomena, such as hesitation or laughter, are natural. Similarly to the conventionally used read aloud speech resources, a bad actor may sound like he is reading aloud, whereas only a good actor can sound sincere and spontaneous (Newell, 2009). Hence, although well acted speech has many similarities to spontaneous speech, good actors are rare whereas spontaneous conversations can be elicited and recorded in large amounts from many different people. Therefore within the work presented here, we will focus entirely on speech from spontaneous conversations, although elicited within the controlled environment of a recording studio.

1.2 Conversational Speech Synthesis

Unit selection and HMM-based speech synthesis frameworks rely on the recorded speech providing phonetic coverage; coverage of the different speech units in relevant contexts, to build high quality synthetic voices. In conventional speech synthesis the speech unit is often based on the phoneme and the contexts include features that affect the phonetic properties of the phoneme, e.g. neighbouring phonemes, position of the phoneme in syllable and utterance, etc. The recorded speech resources then contain read aloud sentences that are pre-selected to provide the desired phonetic coverage. In general, better phonetic coverage gives better quality of the synthetic speech (Clark et al., 2007).

The previous research on speech synthesis with spontaneous or acted speech resources have to a large extent focused on selecting whole dialogue acts (Gustafsson and Sjölander, 2004; Campbell, 2005; Romportl et al., 2010), in particular when the dialogue acts were considered important for regulating the conversation, e.g. backchannels (“Yeah.”, “Too bad.”) or phrases like “Could you repeat that?”, or express affective content e.g. ”Hi how are you?” or “I’m so sorry about that.”. The propositional content was however often synthesised from sub-word units with synthetic voices built from

neutrally read aloud sentences (Romportl et al., 2010; Adell et al., 2010). But, in a conversation many utterances contain both propositional content and speech phenomena such as discourse markers and filled pauses. In the example from our data (described in chapter 3) the propositional content is bold faced and discourse markers and filled pauses are in italics:

"*yeah exactly and even like uh* **I'll go and see bad movies that I know will be bad** *um just* **to see why they're so bad**".

Campbell (2006) describes this as propositional content being "wrapped" in speech phenomena that structure the utterance in the interaction or signal affective meaning. The ability to integrate propositional information with discourse markers and filled pauses in coherent synthetic utterances is therefore an important step towards synthetic voices for believable characters that can express themselves in a manner more similar to human conversation.

The problem of utilising speech from a conversation to build synthetic voices that can synthesise propositional content wrapped in discourse markers and filled pauses is that there is less control over the phonetic coverage in spontaneous speech than in the conventionally used speech resources. Therefore, in order to synthesise high quality speech from spontaneous speech resources one has to a) accept that what can be said with the voice is limited, and for example just select whole phrases, b) develop other synthesis techniques, or c) develop methods to regain control over phonetic coverage, by for example blending speech from different sources. In this work we will consider a number of blending approaches within the unit selection and HMM-based frameworks.

1.3 Perceptual Evaluation

To support or refute the research hypothesis (see section 1.4) we will conduct perceptual evaluations with human participants. Acoustic and linguistic analysis of natural and synthetic speech will also be conducted to provide motivation for the results of the perceptual evaluations.

In conventional speech synthesis, one of the most common evaluation methods is to play isolated utterances of synthetic speech to listeners and let the listeners self-rate perceived naturalness. The listeners' ratings of naturalness have been shown to be negatively affected by the presence of acoustic artefacts associated with synthetic speech errors, such as F0 and spectral discontinuities (Mayo et al., 2005). Evaluating

the naturalness of a synthetic utterance gives information of the overall quality; of propositional content as well as discourse markers and filled pauses. Therefore, we will evaluate the naturalness of our synthetic voices. But naturalness is not enough.

Two synthetic speech samples can be perceived as differing in other aspects than naturalness: the utterances could have different prosodic properties (e.g. fast/slow), different linguistic properties (e.g. casual/formal), or one utterance could sound like it was spoken spontaneously in a conversation and the other like it was read aloud from a newspaper. Previous research that has evaluated other aspects of synthetic speech than perceived naturalness include e.g. evaluations of how "colloquial" (Werner et al., 2006) or "spontaneous" (Lee et al., 2010) an utterance sounds, or whether an utterance has a "joyful","sad", "rough" or "neutral" speaking style (Yamagishi et al., 2005). We will investigate to what extent listeners perceive that synthetic voices built from conversational speech data also exhibit recognisable conversational characteristics, and to what extent this perceived "conversationalness" is distinct from perceived naturalness.

Whereas evaluating a conversational speaking style is intended to capture a general quality, we will also investigate to what extent our conversational speech synthesis can convey specific pragmatic meanings. Listeners' perception of specific phonetic properties of synthetic utterances have been evaluated by requesting participants to listen for them, e.g. by requesting them to listen for placement of hesitation (Carlson et al., 2006) or locate the most prominent word (Strom et al., 2006). Similarly, we will investigate if certain discourse markers and filled pauses affect the perceived meaning of a synthetic utterance so that it conveys certainty or uncertainty.

1.4 Research Questions and Hypothesis

Our objective is to create a synthetic voice which is perceived as both natural and conversational by utilising speech from a spontaneous conversation to build the voice. The research questions that we will address to achieve this objective are:

- How to obtain spontaneous conversations under the controlled conditions required for building high quality synthetic voices.
- How to constrain the rich variety of speech phenomena in a spontaneous conversation to create a controlled dataset of conversational utterances from which we can automatically build high quality synthetic voices in conventional speech synthesis systems.

- To what extent can we alleviate the lack of control over phonetic coverage in spontaneous speech resources by blending conventional pre-selected and neutrally read aloud data with data from a conversation.

- To what extent does the inclusion of conversational speech in synthetic voices influence listeners' impression of conversational speaking style and pragmatic meaning of synthetic utterances.

The hypothesis of this thesis is that incorporating conversational speech into a database of neutrally read aloud speech can add conversational characteristics to an otherwise neutral synthetic voice without causing a negative impact on the perceived naturalness. In contrast, our null hypothesis is that the differences between conversational speech and neutrally read aloud sentences are too big, and the use of speech from a spontaneous conversation in synthetic voices will result in no improvement when synthesising conversational material.

To test this research hypothesis we will build a series of voices constructed with and without conversational data, as well as with and without methods to support appropriate blending of speech data with different speaking styles. We will evaluate these voices in terms of their naturalness and in terms of their conversational speaking style. To conclude we will test that a synthetic voice built with conversational speech can convey pragmatic information, such as certainty and uncertainty.

1.4.1 Structure of the Thesis

The rest of this thesis is outlined as follows: chapter 2 gives the background to conversational speech, unit selection and HMM-based speech synthesis and previous research on conversational speech synthesis. Chapter 3 describes the recording, transcription and analysis of the recorded conversations. Chapter 4 describes the details of building the synthetic voices, and Chapter 5 describes the perceptual evaluations of the synthetic speech. Finally, chapter 6 contains a concluding discussion.

Chapter 2

Background

In this chapter we will start by giving a broad introduction to conversation in section 2.1, before describing in section 2.2 the conversational speech phenomena that are the focus of this thesis. In section 2.3 we will give an overview of the unit selection and HMM-based speech synthesis frameworks that were used to build the synthetic voices in chapter 4. In section 2.4 we will review previous approaches to conversational speech synthesis.

2.1 What is Conversation?

Human face-to-face interaction is recognised as the foundation of human communication in research areas ranging from sociology (Goffman, 1967) to phonetics (Local and Walker, 2005). In the interface between sociology and phonetics we find the everyday conversation (Clark, 1996). We use Clark (1996) to give a broad introduction to conversation, because he takes into account both the private perspective of the individual participants in the conversation as well as the coordination of the participants' individual actions through an observable signal that Clark (1996) refers to as language. Although, face-to-face conversation includes bodily, facial and vocal gestures, it is primarily a linguistic activity (Clark, 1996), and within this thesis we will focus entirely on the speech signal of the conversation.

The core claim about language use in conversation in Clark (1996) is that it is a joint action. Conversation requires coordinated interaction between a speaker and an addressee. What needs to be coordinated is what the speaker means and what the addressee understands about the speaker's intended meaning. A short example of participants expressing meaning and understanding in a conversation from Clark (1996,

p.227) is shown below[1]:

Roger: *now, - um do you and your husband have a j- car*

Nina: *- have a car?*

Roger: *yeah*

Nina: *no -*

The example from Roger's and Nina's conversation is analysed by Clark (1996) as follows: Roger believes that he has expressed his meaning to Nina with his first turn. Nina confirms that she believes that she has understood Roger, except for the last part which she thinks was *have a car*. Roger then concludes that if he confirms that he meant *have a car* then Nina will have understood what he meant, so he says *yeah*. Nina then confirms that she has understood what Roger meant by his first question by answering it with *no*.

2.1.1 Structure of Conversation

The individual contributions in the conversation to the shared social context or discourse, e.g. Nina's *-have a car?*, have been categorised and analysed in the research literature as e.g. speech acts (Searle, 1969), turns (Sacks et al., 1974), and more recent derivations of turns or speech acts as dialogue acts in speech synthesis and spoken dialogue systems (Campbell, 2005; Traum et al., 2008; Bunt et al., 2010). The main difference between turns and dialogue acts is that turns focus on the process of coordinating who speaks when (turn-taking) in the conversation (Sacks et al., 1974), and dialogue acts focus on the pragmatic function of what was said (Bunt et al., 2010).

Part of the definition of a dialogue act in Bunt et al. (2010) is the requirement of at least two participants: a speaker and an addressee. Clark (1996) argued that dialogue act type is negotiated between speaker and addressee and depend on both the addressee's understanding and acceptance of the speaker's meaning, and the speaker's acceptance of the addressee's understanding. Clark (1996) gives an example where the utterance *"Sit here"* can be interpreted as an order, a request, an offer or an advisory, and it takes both speaker and addressee to negotiate which one it will be. For example, by replying *"Yes, sir."* the addressee signals that (s)he understands and accepts *"Sit*

[1]We present transcribed examples from other work with original annotations. In these examples, hyphen or dot are used to denote silences, colon is used to denote prolongation of a segment, and other punctuation marks are used to denote phrase endings, e.g. question mark is denoting a question.

here" as an order, whereas if the addressee replies with *"What a good idea!"* the addressee signals that (s)he understands and accepts *"Sit here"* as an advisory. The recording, transcript or dialogue act annotation of a conversation only displays what did happen, not which other options were available and considered. The structure of a conversation can appear pre-determined when analysed after the fact, but is the result of locally negotiated contributions (Clark, 1996).

We will use the terms turn and dialogue act when reporting previous research that have used them. The speech data analysis (see chapter 3) for the synthetic voices described in chapter 4 of this thesis, did not use turns or dialogue act annotation (see section 2.2). Therefore we will often use the more neutral term utterance when referring to a delimited stretch of speech.

2.2 Conversational Speech for Speech Synthesis

The problem of synthesising conversation consists of generating appropriate speech at an appropriate time in an interactive setting. That problem can be divided into two parts; the interactive part and the static part. The interactive part is addressed in research such as Traum et al. (2008), where their animated characters engage in limited conversations. One of the limiting factors of these characters is the lacking ability to generate speech with conversational characteristics beyond a limited set of pre-recorded prompts. Generating speech with conversational characteristics in utterances that are not pre-recorded represents the static part of the conversational speech synthesis problem. In the static view of conversation the features of recorded conversations are analysed and duplicated. This is the approach taken in this thesis. We concentrate on the description and analysis of recorded conversational data and try to duplicate it. To evaluate this approach we then carry out a limited "interactive" experiment where we present single sentences to our subjects to see if the intended function of conversation has been successfully synthesised; in this case giving an impression of certainty or uncertainty (see section 5.6). As a starting point for this approach we will consider one of the most common features of conversational speech, so called "wrappers" (Campbell, 2006).

The topics and participants of conversations vary from occasion to occasion, but many of the speech phenomena in conversation are recurring across different topics and different speakers. These recurring phenomena are the key to the unit selection and HMM-based speech synthesis frameworks' ability to synthesise utterances that

are not pre-recorded. In section 1.2 we gave an example of conversational speech where the content was separated into wrappers and propositional content. In a corpus of 150,000 utterances from one person's everyday conversations Campbell (2006) found that about half of the utterances consisted only of these recurring wrappers. Campbell (2006) argued that the wrappers were used in conversation to regulate the flow (e.g. turn-taking), express inter-personal relationship (e.g. formal/informal) and express affective content (e.g. agreement, disagreement or hesitation). These wrappers are generally not well synthesised with conventional synthetic voices. Improving synthesis of the wrappers would make synthetic voices more suitable for applications such as believable characters (Traum et al., 2008; Romportl et al., 2010).

Based on the research literature we divided the wrapper category into discourse markers, filled pauses and backchannels. We will focus on discourse markers and filled pauses, because the challenges for synthesising backchannels are different than the challenges for synthesising utterances with discourse markers and filled pauses. Discourse markers and filled pauses mainly occur together with propositional content in longer utterances (Schiffrin, 1987; Clark and Fox Tree, 2002), whereas backchannels are often isolated word tokens (Hockey, 1993; Gravano et al., 2007). Given a database of conversational speech, as in chapter 3, many backchannels already exists in the recordings and the challenge would be to time them appropriately in a conversation. The challenge that we will address is to integrate discourse markers and filled pauses with propositional content to synthesise utterances that are not pre-recorded.

In sections 2.2.1, 2.2.2 and 2.2.3 we will describe phonetic properties and pragmatic functions of the discourse markers, filled pauses and backchannels. The majority of the reviewed research has analysed the phonetic properties with respect to manually labeled discourse features and/or pragmatic functions, e.g. the effect of the preceding utterance on listeners classifications of *okay* tokens (Gravano et al., 2007), or the differences in $F0$ trajectory of *okay* when used as backchannel or discourse marker (Hockey, 1993). In our approach we will investigate the use of the lower level features that are automatically extracted in our speech synthesis systems, such as phoneme sequence and utterance position. The decision to use automatically extracted low level features does have limitations in representing and synthesising meaning contrasts of conversational speech phenomena. But, the main motivations behind our bottom-up approach to conversational speech synthesis were:

- The phonetic content of recorded speech is fixed, whereas the pragmatic function of synthetic speech will be interpreted in a new discourse (see section 2.1.1).

Therefore, it was considered more important that the extracted features supported synthetic speech without acoustic artefacts, rather than that the features reflected every meaning contrast in a recorded conversation. The evaluations of the synthetic voices in sections 5.2 and 5.5 show that the approach was successful in synthesising more natural conversational style utterances than conventional synthetic voices.

- Discourse markers and filled pauses consist of a limited set of word tokens that are frequently occurring at or around phrase boundaries. Many important characteristics of discourse markers and filled pauses can therefore be identified for speech synthesis through low level features. The claim will be substantiated by: a) showing preserved phonetic properties of discourse markers and filled pauses in synthetic speech (in section 4.2.6), and b) showing the effect of discourse markers and filled pauses in synthetic speech on listeners' perception of pragmatic function (in section 5.6).

Ambiguous examples where low level features are insufficient can be constructed. For example, does stand-alone *right* mean the opposite of *left*, or is it a backchannel? In our recorded conversation in chapter 3 there are 167 stand-alone *right* (see table 3.4), and all of them are backchannels. Thus, the low level features often capture a token's prototypical function and associated phonetic properties.

2.2.1 Discourse Markers

Discourse markers include mainly words and expressions that are frequent in conversations, such as: *actually*, *basically*, *because*, (examples from Hirschberg and Litman, 1993), *oh*, *well*, *but*, *you know*, *I mean* (examples from Schiffrin, 1987). Different authors have used different terms to refer to similar sets of words and phrases, e.g. cue phrases (Grosz and Sidner, 1986), editing terms (Levelt, 1983), lexical fillers (Lickley, 1994). In this thesis we will use the term discourse markers (Schiffrin, 1987).

We will focus on describing discourse markers that: a) were used by the speaker in the data described in chapter 3, and b) have been analysed with respect to their phonetic properties and pragmatic functions. The literature review shows that different discourse markers are often associated with one or a few prototypical pragmatic functions associated with their local phonetic properties. The "lexical form" is an important part of this local phonetic context and Jurafsky et al. (1998) treated *yeah*, *oh yeah*, *yeah (LAUGH)* and *well yeah* as separate types. Our method of extracting low level

features from conversational speech that contains a rich variety of discourse marker types, should therefore identify the local phonetic properties of different discourse markers and thereby support synthesis of different discourse markers in a manner that sounds natural.

2.2.1.1 Yeah and Okay

The most frequent word in our data, described in chapter 3, was *yeah*. That *yeah* is frequent in conversations is also the consensus in the literature (Jurafsky et al., 1998; Fuller, 2003; Benus et al., 2007). Frequent pragmatic functions of *yeah* include backchanneling (see section 2.2.3), yes-answer, and agreement (Jurafsky et al., 1998).

Gravano et al. (2007) classified tokens of *okay* from 12 task-oriented conversations into pre-defined categories including: backchannel, agreement and discourse marker at beginning of turn. A subset of the tokens, matched with respect to labeler agreement (full to none), were selected for a perception task, where participants were asked to assign one of the three categories to each token. The different tokens of *okay* were played both in isolation, and in the context of previous and current turn. An example is shown below, where the *okay* in boldface was the token to be classified:

Speaker A: yeah - um there's like there's some space there's

Speaker B: **okay** - I think I got it

Segmental, prosodic and discourse features were extracted from the *okay* tokens to analyse which were correlated with the participants' classifications. An important finding was that different features were correlated with the participants' classifications when *okay* was played in isolation or in its context. In isolation there were correlations with the segmental quality and duration of phonemes, but in context the strongest correlations were related to duration of silence between turns and the length of speaker B's turn (where the *okay* to be classified was). Both in isolation and in context pitch contour showed relatively strong correlation with classifications, where a rising contour was correlated with backchannels and a falling contour was correlated with discourse markers (Gravano et al., 2007).

The relation between pitch contour and discourse function for *okay* (backchannel or discourse marker) was previously established also in Hockey (1993). A falling pitch contour was associated with a function as discourse marker, and a rising pitch contour was associated with a function as backchannel (Hockey, 1993). In addition, there

were also *okay* tokens with a flat contour, but no pattern related to function could be established across speakers, and Hockey (1993) suggested that pitch contour alone was not a sufficient cue to identify discourse function.

2.2.1.2 I Mean and You Know

In the analysis by Schiffrin (1987), *I mean* signalled a speaker's orientation towards and modification of their own speech. This analysis is in accordance with the function of *I mean* also in Levelt (1983). As in the example below where *I mean* signalled a change from *I don't know* to *I know* (Schiffrin, 1987, p. 301):

> But oh I don't know the rabb- **I mean** I know him, but I'm- I- not actively, as far as I'm concerned

You know, often pronounced (in Schiffrin's notation) as *y'know*, was used to refer to shared knowledge of the speaker and hearer, where a rising intonation signalled more uncertainty about the shared knowledge than a falling intonation (Schiffrin, 1987). *You know* could also be used to elicit confirmation from the hearer, as in the example adapted from Schiffrin (1987, p. 292):

> Irene: [...] he had taken over the synagogue, which remained there:**y'know**?
>
> Sally: Yeh, I remember.

2.2.1.3 And and But

The discourse connective *and* was used to coordinate and continue actions, e.g. signal relation to previous turns. Whereas the discourse connective *but* was used to signal contrast, and also disagreement (Schiffrin, 1987). Although it was not analysed in Schiffrin (1987) the collocation *and* followed by a filled pause (*and uh/um*) was relatively frequent in her examples. Local (2007) showed that *and* in *and uh/um* had very consistent phonetic properties compared to when *and* as discourse marker was not followed by a filled pause. Local (2007) argued that whereas *and* was used to continue the current topic, *and uh/um* was used to return to a prior topic.

2.2.1.4 So

So can be used to signal turn transitions, as in the example adapted from Schiffrin (1987, p. 219):

Henry: [...]

Henry: **So**:uh...but we buy beer and...cake and that's-we spend it out of our own money.

Henry: **So**:eh:

Debby: **So**, when Henry's gone, what do you do?

Where, according to the analysis, Henry offered the turn with the first *so*, but then continued when nobody took it, and offered the turn again with the second *so*, at which point Debby took the opportunity and asked the third participant a question (Schiffrin, 1987).

An analysis of stand alone *so* in American English conversations showed how phonetic properties differed with respect to discourse function (Local and Walker, 2005). The comparison was made on two types of stand alone *so*: "holding-so" and "trailoff-so". The "trailoff-so" was a signal to the conversational partner that the previous topic was finished and that the partner was welcome to take the turn and initiate a new topic. The "holding-so" on the other hand signalled that the speaker had not finished the current topic and therefore continued speaking after the silence, without the conversational partner attempting to take the turn.

All instances of "holding-so" and "trailoff-so" were in the immediate phonetic context surrounded by silence, and they had a variety of phonetic properties with respect to: vowel quality, duration, pitch contour and voice quality. But the phonetic differences with respect to discourse function was that "holding-so" was significantly louder, had higher f0 and was less creaky than "trailoff-so" (Local and Walker, 2005).

2.2.1.5 Discourse Marker Summary

Discourse markers consist of frequent words and expressions that are used to express a wide range of functions in conversation (Schiffrin, 1987). The different discourse markers are often associated with a few prototypical functions as described in sections 2.2.1.1-2.2.1.4. For example, expressing agreement with *yeah*, signalling relation to a previous topic with *and uh* or asking for confirmation with *you know*.

In our approach, outlined in section 2.2, the phonetic properties and functions of the different discourse markers were represented for speech synthesis through shallow linguistic features. If we review the findings we have presented for discourse markers in sections 2.2.1.1-2.2.1.4 we find that a majority can be modelled to a large extent

function	propositional	initial discourse marker	final discourse marker
tokens	... *nowadays* **you know** *automatically who's gonna die...*	**sil you know** *I was...*	*...with the helmet sp* **you know sil**
function	propositional	backchannel	initial discourse marker
tokens	... *my* **right** *foot ...*	**sil right sil**	**sil right** *sp well you guys...*
function	backchannel	confirmation	initial discourse marker
tokens	**sil yeah sil**	**yeah yeah yeah**	**sil yeah** *I know it's...*
function	propositional	hesitation	more hesitation
tokens	... *it's fast ...*	... **it's uh** *it's blue ...*	... **but uh** *I think* **it's uh** *it's cool*

Table 2.1: Examples of how shallow linguistic features, such as representation of utterance position and word/phoneme context, can distinguish between the functions of different tokens. The *sil* represents utterance beginning or end. The *sp* represents utterance internal pauses. Isolated utterances with *yeah* or *right* are often backchannels, written *sil yeah sil* or *sil right sil* in this table. In the beginning or end of utterances *yeah* and *you know* are often discourse markers, e.g. in utterances starting with *sil yeah I...* or ending with *...sp you know sil*. Filled pauses in an utterance signal hesitation, e.g. *...it's uh it's blue* whereas only propositional content often does not signal hesitation, e.g. *...it's fast...*. Orthographic representation of these tokens, e.g. *yeah*, *right* or *uh*, together with their immediately surrounding context was therefore expected to be sufficient to represent their phonetic properties for speech synthesis. The examples are taken from the conversation with Johnny in chapter 3.

through immediate phonetic and word context. Table 2.1 exemplifies how features such as phoneme sequence and phrase position can capture prototypical function distinctions.

2.2.2 Filled Pauses and Hesitation

The term filled pause was coined in Maclay and Osgood (1959) (reprinted in Jakobovits and Miron (1967)) as a contrast to unfilled pauses (silence or phoneme prolongation) in an analysis of hesitation phenomena in English spontaneous speech. Filled pauses are sometimes classified as disfluencies, but they have linguistic properties more in common with other "filler" items, e.g. *I mean* (Levelt, 1983). The transliteration of English filled pauses differs within the literature, but in this thesis we will use *um* and

uh.

Clark and Fox Tree (2002) argued that filled pauses should be considered normal English words that signal delay in speech, with slightly different meanings of *um* and *uh*. The postulated meaning difference was argued against in O'Connel and Kowal (2005), but for this thesis the word-like properties of filled pauses are the focus of attention, not their meaning difference.

2.2.2.1 Linguistic Properties

As we mentioned in section 2.2.2, filled pauses are sometimes analysed together with disfluencies such as repetitions, and in these analyses filled pauses were found to be a very frequent (if not the most frequent) disfluency type (Shriberg, 1996; Lickley, 2001). The frequency and type (*um* or *uh*) of filled pauses are to a large extent individual. In the analysis by Clark and Fox Tree (2002) of the London-Lund corpus, speaker's filled pause rate varied between 1% and 9% of the total number of word tokens and some speakers showed a clear preference for either *um* or *uh*, but averaged over all speakers they were used about 50% of the time each.

The majority of filled pauses occurred at syntactic boundaries, or after the first word and less frequently in other positions (Clark and Fox Tree, 2002). The rate of filled pauses also varied with dialogue act type, with more filled pauses in replies to wh-questions, instructions and negative answers, than in y/n-questions or positive replies (Lickley, 2001).

Shriberg and Stolcke (1996) showed that utterances that contained repetitions or filled pauses had significantly lower bigram and trigram transition probabilities than fluent utterances.

2.2.2.2 Phonetic Properties

Although filled pauses are word-like, their specific phonetic properties differentiate them from other words. In this section we will describe the phonetic properties of filled pauses that have been reported in the research literature.

Filled pauses consisted of a steady vowel part that was sometimes followed by an /m/ (O'Shaughnessy, 1992). The vowel quality of filled pauses was often close to a schwa, but could also have other vowel qualities (Shriberg, 1999). But one of the most distinguishing characteristics of filled pauses was their duration.

Shriberg (1999) reported a median duration for filled pauses of approximately

300ms, but with a large variation of duration (from about 50ms up to almost a second). The duration differed to some extent both between *um* and *uh*, where *um* was on average 60-100ms longer than *uh* (Brennan and Williams, 1995; Fox Tree, 2001), and between filled pauses at syntactic boundaries (200-500ms) and within clauses (170-320ms) (O'Shaughnessy, 1992).

Filled pauses generally had a lower F0 than the rest of an utterance, but filled pauses at syntactic boundaries tended to have a higher F0 onset than clause internal ones (O'Shaughnessy, 1992). The pitch contour of a filled pause can be falling, level or rising (Clark and Fox Tree, 2002). Shriberg and Lickley (1993) showed that the F0 of clause internal filled pauses correlated with F0 values of surrounding F0 peaks (e.g pitch accents), regardless of if the filled pause was separated from the surrounding speech with a silent pause.

Filled pauses were sometimes cliticised onto prior words so that e.g. *and uh* or *but um* were pronounced as *an duh* and *bu tum* (Clark and Fox Tree, 2002). As a hesitation phenomenon filled pauses are often associated with a prolongation of at least the preceding syllable, but the reported evidence for this particular phenomenon is sparse, and the only explicit support we have found comes from Adell et al. (2008). Other research has analysed hesitation prolongation as a more general phenomenon preceding disfluencies, such as repetitions and filled pauses, that also included usage of fully pronounced versions of e.g. *a*, *the* or *to* (Shriberg, 1999).

2.2.2.3 Pragmatic Functions of Filled Pauses

The reason to synthesise filled pauses and other conversational speech phenomena in a natural manner, is to communicate something to the listeners. Psycholinguistic studies have shown how a speaker's use of filled pauses affect the listeners in various ways.

Brennan and Williams (1995) showed that listeners' impressions of a speaker's certainty of an answer was affected by the presence of filled pauses. Corley et al. (2007) showed that listeners experienced fewer problems of integrating unpredictable words into their context when they were preceded by a filled pause. Numerous other psycholinguistic studies (e.g. Arnold et al., 2007) have shown that the listeners' attention was directed towards discourse new referents when there was a filled pause before a referent. Arnold et al. (2007) also showed that this effect was cancelled when listeners were told that the speaker suffered from agnosia, an inability to recognise or name objects, showing that listeners took into account why the speaker hesitated. But it is also worth mentioning that Corley et al. (2007) and Arnold et al. (2007) used (the

authors') acted and not actual spontaneous hesitations, recorded in carrier sentences, such as in the example from (Arnold et al., 2007, p. 916): *"click on thee uh ..."*. Those acted hesitations might be more prominent than in actual spontaneous speech, because Lickley (1995) found that people failed to consciously detect approximately half of the utterance internal filled pauses in spontaneous speech.

2.2.2.4 Filled Pause Summary

The specific phonetic properties of filled pauses described in section 2.2.2.2 are different from the properties of other speech phenomena (see e.g. Adell et al. (2010)). Conventional unit selection and HMM-based synthetic voices will therefore not generate filled pauses with natural phonetic properties, unless there are special solutions as in Adell et al. (2010).

The filled pauses are in this thesis written as *um* or *uh*. This is also how they are represented in the pronunciation lexicons of our speech synthesis systems, together with their phoneme sequences. This representation differentiates the filled pauses from other words, and because filled pauses exist in our speech data their phonetic properties are well captured through our bottom-up approach outlined in section 2.2 of utilising phoneme sequence, utterance position and other shallow features to synthesise filled pauses. Table 2.1 exemplifies how these features in the typical case capture hesitation or uncertainty about the propositional content through the presence or absence of filled pauses in the text.

In summary, speakers use filled pauses when hesitating, and listeners, to some extent, recognise and interpret the reason for the speaker's hesitation. In section 5.6 we will investigate the contribution of filled pauses on the perception of (un)certainty in synthetic speech.

2.2.3 Backchannels

Backchannels are signals that the listener is involved in the conversation, but does not want to take the turn from the speaker (Gravano et al., 2007). Backchannels often have the same lexical realisations as discourse markers, e.g. *okay*, *yeah*, but some tokens, e.g. *uh-huh*, have a purely backchannel function (Hockey, 1993). The phonetic properties, such as pitch slope, have been found to differ between *okay* tokens classified as backchannels or discourse markers. Another important classification cue was that the backchannels were isolated from speech by the same speaker with silent pauses

Hockey (1993); Benus et al. (2007); Gravano et al. (2007).

The speech data in chapter 3 that was used to build the synthetic voices in chapter 4 contained backchannels. These backchannels could, in unit selection, be selected as pre-recorded prompts based on orthographic content and phrasal context (see table 2.1). The timing of backchannels in conversation is, in contrast, a major challenge. Examples of work that have focused on the timing of backchannels include Schröder et al. (2008) and Romportl et al. (2010). These systems typically use a full dialogue system and an embodied conversational agent (Schröder et al., 2008; Romportl et al., 2010). Our work did not require modelling timing in conversation, and synthesis of backchannels will therefore not be considered further in this thesis.

2.3 Speech Synthesis

Unit selection and HMM-based speech synthesis are currently the two dominating frameworks in speech synthesis. They both utilise recordings of speech to build synthetic voices that capture the characteristics of the speech and speaker in the original recordings and enable synthesis of utterances that are not pre-recorded.

The unit selection and HMM-based speech synthesis frameworks are based around the same assumptions about speech as a sequence of context-dependent sub-word speech units. For English, the sub-word speech unit is generally the phoneme, and the context includes features that affect the phonetic properties of the phoneme, e.g. neighbouring phonemes, syllable position, utterance position or prosodic prominence. The different engineering solutions of unit selection and HMM-based speech synthesis have certain consequences for the resulting synthetic speech. In unit selection the phonetic detail of the original speech recording is preserved, but the concatenation of sub-word units in connected speech can result in audible acoustic artefacts at concatenation points. In HMM-based speech synthesis the speech is vocoded which results in a degradation of speech quality, and the training and generation schemes result in a loss of some of the original phonetic detail, but the training and generation schemes also result in more consistent speech quality than unit selection. In this thesis we will investigate whether unit selection is *robust enough* to make good quality synthetic voices from conversational speech, and whether HMM-based speech synthesis is *sensitive enough* to preserve important phonetic detail of conversational speech phenomena.

We used three different systems to build the synthetic voices in chapter 4: the CereVoice (Aylett and Pidcock, 2007) unit selection system, and the speaker-dependent

(Zen et al., 2007) and speaker adaptive (Yamagishi et al., 2009) HMM-based speech synthesis systems. These three systems have been shown to synthesise good quality speech from conventional speech resources of read aloud sentences, in for example the Blizzard Challenge (Andersson et al., 2008; Karaiskos et al., 2008; Yamagishi et al., 2008). Additionally, techniques have been developed for these systems to synthesise different "emotions" from recordings of more expressive speech (Aylett and Pidcock, 2007; Yamagishi et al., 2005, 2004). The systems were therefore considered adequate candidates for the challenging task of utilising conversational speech to build natural-sounding synthetic voices.

In section 2.2 we motivated how frequent conversational speech phenomena could be represented for speech synthesis through shallow linguistic and phonetic features. This analysis can be automatically made by our speech synthesis systems. But, unit selection and HMM-based speech synthesis require phonetic coverage (see section 2.3.1) in order to build high quality voices. The lack of control over phonetic coverage in conversational speech led us to investigate "blending" of conversational and read aloud data in the synthetic voices. Section 3.5.1 contains an analysis of the phonetic coverage in the conversational and read aloud data. The purpose of the blending was to use the read aloud data to boost the phonetic coverage, and thereby allow high quality synthetic speech, while maintaining the conversational characteristics from the conversational speech data. In order to avoid it being obvious to listeners that the synthetic voices were built from two different sources of data, and therefore sound less natural, the developed blending techniques needed to take into account the phonetic differences between the conversational and read aloud speech data. A comparison of the general phonetic properties of our recorded speech is shown in section 3.5.2.

Conventional speech synthesis evaluations, as in the yearly Blizzard Challenge workshop (Black and Tokuda, 2005; King and Karaiskos, 2010), generally focus on evaluating the naturalness and intelligibility of synthetic speech. Naturalness is evaluated by letting listeners self rate the perceived naturalness of synthetic speech. Intelligibility is evaluated by letting listeners write down the perceived orthographic word sequence of a synthetic utterance. General differences other than naturalness have often been evaluated in the research literature as a difference in speaking style, e.g. which utterance sounds more *"joyful"*, *"sad"*, or *rough* (Yamagishi et al., 2005), or which utterance sounds more *"colloquial"* (Werner et al., 2006) or *"spontaneous"* (Lee et al., 2010). More local properties of synthetic utterances have been evaluated by requesting listeners to locate e.g. hesitation in the beginning, mid or end of a synthetic utterance

(Carlson et al., 2006) or identify the prosodically most prominent word in an utterance (Strom et al., 2006). We will follow the conventional evaluation paradigm, to contrast the impact of conversational data on naturalness and speaking style of our conversational synthetic voices compared to conventional voices built from read aloud sentences. We will also evaluate whether a conversational voice communicates pragmatic information more efficiently than a conventional voice. The details of the perceptual evaluations are described in chapter 5.

In sections 2.3.2 and 2.3.3 we will describe our unit selection and HMM-based speech synthesis systems. But first we will give an overview in section 2.3.1 of the carefully read aloud isolated sentences that are generally used for speech synthesis.

2.3.1 Conventional Speech Resources

Conventional unit selection and HMM-based speech synthesis systems rely on recordings of read aloud isolated sentences that are selected to provide phonetic coverage. Phonetic coverage for synthesis means that the speech unit should be present in all relevant segmental and prosodic contexts. In particular, the contexts should cover the intended target domain or text genre (Clark et al., 2007). In this thesis the speech unit in the unit selection system is the diphone (Aylett and Pidcock, 2007), and in the HMM-based speech synthesis system the quinphone (Zen et al., 2007). The diphone stretches from the middle of a phoneme to the middle of the next phoneme. This facilitates concatenation of units in unit selection since the phonetic properties are more consistent across contexts in the middle of the phoneme (Clark et al., 2007). The quinphone is an extension of the triphone used in speech recognition and stretches from the beginning of the first phoneme to the end of the fifth phoneme (Young et al., 2006).

In Clark et al. (2007) the CMU Arctic database (Kominek and Black, 2004) was considered to give a minimum phonetic coverage. The Arctic database consists of approximately 1200 sentences, 5-15 words long, collected from fiction. The Arctic database contains at least one of about 90% of the possible diphones in their lexicon, when only lexical stress was considered as phonetic context (Kominek and Black, 2004). Richer phonetic coverage generally includes sentences from a variety of text genres, as in the data used for the Blizzard Challenge 2008 (Karaiskos et al., 2008) which contains about 8000 sentences from e.g. news, fiction and addresses that were originally recorded by Strom et al. (2007, 2006). This data was collected to obtain coverage of phrase boundaries and pitch accents, in addition to lexical stress (Strom

et al., 2007, 2006). In general, better phonetic coverage results in better synthetic speech quality; for both unit selection and HMM-based synthetic voices (Aylett and Yamagishi, 2008).

According to Clark et al. (2007) there are two problems with the need for phonetic coverage in speech synthesis:

- The need for phonetic coverage quickly increases the number of needed sentences as more prosodic contexts are considered.
- It is difficult to consistently record a large amount of speech from a single speaker over multiple sessions.

2.3.1.1 Segmenting Read Aloud Speech

To build synthetic voices from recordings of speech, conventional unit selection and HMM-based speech synthesis systems require that the speech is segmented into a phoneme sequence. The phoneme sequence is typically derived from a forced alignment of an orthographic transcription to the speech signal. Poorly segmented speech results in poor synthetic speech quality. The alternative of manual segmentation was rejected on the basis that it is too resource intensive. Thus, investigating to what extent spontaneous speech can be automatically processed for synthesis, compared to the conventional read aloud sentences, is a key problem. Therefore, we utilised forced alignment to derive a phoneme sequence also for the carefully transcribed conversational utterances described in section 3.4.

Speech segmentation for synthesis consists of two problems: determining the phoneme sequence and aligning that sequence to the speech signal. The HTS system does not include the tools for determining the phoneme sequence and it was in this thesis determined with the Festival or CereVoice text processing modules. These modules are often termed the *front-end*. The front-end converts transcriptions to phoneme sequences using pronunciation lexicons, phrasing rules and other phonological rules. The pronunciation lexicon lists valid phonemic pronunciations (generally citation form pronunciations) of isolated words, thereby simplifying the problem of determining the phoneme sequence. Forced alignment, as outlined in Young et al. (2006) and implemented in a similar manner in the Festival and CereVoice systems, generally provides accurate alignment of carefully read aloud sentences. The forced alignment modules in both Festival and CereVoice are implemented using the HTK toolkit (Young et al., 2006). Each phoneme is represented as a three state left-to-right hidden Markov model

(HMM). The HMM phoneme sequence is initially aligned to the speech signal with a uniformly distributed duration. The HMM parameters are initialised with the global mean and variance from the spectral features of all the utterances. Then the HMMs are trained with the Baum-Welch algorithm to find a more accurate alignment. To make additional improvements to the alignment optional silences can be inserted between words during the training, the phoneme HMMs may have multiple Gaussian mixtures to account for some of the phonemic variation in connected speech (phrase position, consonant clusters, etc.) and some pronunciation variation is allowed for, in particular, function words. For example, in the CereVoice system, from the general American lexicon, *and* can be pronounced fully /ænd/ or reduced /ən/, *but* can be pronounced fully /bʌt/ or reduced /bət/, and *the* can be pronounced fully /ðiː/ or reduced /ðə/. The result of the forced alignment is to a large extent dependent on how well the phoneme sequence matches the audio. The listed pronunciation variants allow more variation that make the phoneme sequence a more likely match to the more casual pronunciations in spontaneous speech (see e.g. Nakamura et al., 2008; Aylett and Turk, 2006; Johnson, 2004).

Segmenting conversational speech presents a substantial challenge even when the task is facilitated by having an orthographic transcription of the audio. The challenges compared to carefully read aloud sentences are that spontaneous speech contains laughter and other non-speech sounds and it contains more word fragments, mispronunciations, phoneme elisions and reductions.

2.3.2 Unit Selection

The CereVoice diphone unit selection speech synthesis system was developed by CereProc Ltd and is available for academic and commercial use (Aylett and Pidcock, 2007).

The CereVoice synthesis engine is based around the concept of a “spurt” of speech which is defined as the speech between two silent pauses. An input text to be synthesised must first be converted into spurt-sized XML representations. The spurt XML is converted to a target phoneme sequence through look-up in a pronunciation lexicon and applying rules to disambiguate homographs and specify pronunciation reduction variants of function words (Aylett and Pidcock, 2007).

The selection of units in CereVoice follows the general unit selection framework outlined in Hunt and Black (1996). Given a target sentence and a database of speech, the space of heuristically weighted linguistic (target) and acoustic (join) features in the

database is searched for an optimal sequence of diphone-sized units to concatenate into the target utterance. In order to speed up the search, pre-pruning of candidate units is performed before the Viterbi search (Aylett and Pidcock, 2007).

2.3.2.1 Synthesising Different Speaking Styles

The CereVoice system offers the ability to synthesise speech with different speaking styles with the same voice, and in Aylett and Pidcock (2007) this was utilised to synthesise subtle emotions. To realise different speaking styles, subsets of the speech data with different speaking styles were marked with a genre tag. When a specific genre was requested at synthesis time, units from other genres were pruned out before the Viterbi search. If there was insufficient phonetic coverage from the requested genre, units from other genres were included in the Viterbi search.

The genre biasing technique had a large impact on which units were selected (Aylett and Pidcock, 2007). This will be utilised in this thesis both to bias selection towards conversational units and to blend conversational and read aloud speech when there is an insufficient amount of appropriate conversational units.

2.3.2.2 Challenges for Conversational Unit Selection

The conventional speech resources in section 2.3.1 are selected to provide phonetic coverage, because better phonetic coverage gives better quality synthetic speech. In a spontaneous speech resource there is less control over the content, which makes it problematic to achieve phonetic coverage. Therefore we attempted to blend read aloud and spontaneous speech to alleviate the lack of phonetic coverage in our recorded conversation.

The problem with blending is that people can often hear the difference between someone speaking spontaneously or reading aloud (Blaauw, 1992, 1994; Laan, 1997). But, whereas the ability to differentiate between spontaneous and read aloud speech is high for whole utterances, it decreases to chance level for unstressed syllables (Blaauw, 1992). This suggested that some seamless blending of read aloud and spontaneous speech would be possible.

As stated in section 1.4, our objective is to synthesise speech which is perceived as both natural and conversational. The use of speech directly from a spontaneous conversation in the synthetic utterances is likely to preserve a conversational quality to the listeners. But, the blending and segmentation may result in low quality syn-

thetic speech with audible acoustic artefacts. Additionally, if there are too many read aloud units selected in an utterance, it may sound natural, but it may no longer convey any conversational quality. The challenge of blending is therefore to find the trade-off between selecting conversational units to convey a conversational quality to the listeners, and selecting read aloud units to maintain naturalness when there is a gap in the conversational coverage.

2.3.3 HMM-based Speech Synthesis

The speaker-dependent HTS system is an integrated statistical framework based around the hidden Markov model (HMM) for building synthetic voices from recordings of speech (Zen et al., 2007). The general work flow of the HTS system consists of:

- extracting acoustic parameters from speech
- generating context-dependent phoneme representations
- training HMM-based models of acoustic properties for the context-dependent phonemes
- generating speech parameters from the trained models

The training and generation steps in Zen et al. (2007) are described in more detail in sections 2.3.3.1, 2.3.3.2, 2.3.3.3 and 2.3.3.4.

2.3.3.1 Context-dependent Phonemes

The context-dependent phoneme representations define the language related segmental and prosodic categories and dependencies in speech, for both the training and generation parts of HMM-based speech synthesis. The context-dependent phoneme representations are generated from text analysis of the transcribed speech. The text analysis is not part of the HTS system itself and was in this thesis made by the CereVoice system (Aylett and Pidcock, 2007) for the voices in sections 4.2.4 and 4.2.5, and with the Festival system (Clark et al., 2007) for the pilot HTS voice in section 4.2.3.

The context specification for neutral read aloud English is generally similar to Tokuda et al. (2002) or its more recent variants in Zen et al. (2004a) and Yamagishi et al. (2007). The basic speech unit in HTS is the phoneme (it does not have to be, but it is the most commonly used). The context extends all the way from neighbouring phonemes to syllable, word, phrase and utterance level. To model the phonemes' acoustic properties in different segmental and prosodic contexts, the text

is converted into context-dependent phoneme definitions that determine the language-dependent categories of the speech: the phonemes, linguistic and prosodic information such as boundary tones, pitch accents, part-of-speech, etc. Examples of used contexts in Tokuda et al. (2002) are:

- {preceding, current, succeeding} phoneme
- which vowel in current syllable
- position of current phoneme in syllable, word and phrase
- position of current syllable in word, phrase and utterance
- position of current word in current phrase
- stress and accent of {preceding, current, succeeding} syllable
- number of {preceding, succeeding} stressed or accented syllables
- part-of-speech of {preceding, current, succeeding} word
- end tone of current phrase

In section 2.2 we argued that such low level features would suffice to capture important characteristics also for discourse markers and filled pauses.

In (Yamagishi et al., 2005) an additional context: speaking style, was sufficient to blend and preserve different "emotional" speaking styles.

2.3.3.2 Acoustic Analysis

In this thesis, as well as in Zen et al. (2007), we used the STRAIGHT speech vocoder (Kawahara et al., 1999). Excitation and spectral parameters, including their delta and delta-delta, are extracted from the acoustic speech signal as 39 STRAIGHT mel-cepstrals, aperiodicity and $logF0$ (Zen et al., 2007). Additionally, a measure ("global variance") of the variation of mel-cepstral, aperiodicity and F0 per utterance is extracted (Toda and Tokuda, 2007).

2.3.3.3 Excitation, Spectral and Duration Training

In the training phase the acoustic parameters and the context dependent phonemes are jointly trained in an integrated HMM-based statistical framework to estimate Gaussian distributions of duration, excitation and spectral parameters for the context-dependent phonemes (Zen et al., 2007).

To enable simultaneous modelling of voiced and unvoiced sequences of speech, and allow better modelling of phoneme duration the basic hidden Markov model have

been extended into a multi-space probability distribution hidden semi-Markov model (MSD-HSMM) that is used for both training and generation (Zen et al., 2007).

The context-dependent phonemes (see section 2.3.3.1) result in a very large number of model definitions with very few instances of each unique context-dependent phoneme. It is not feasible to get training data that covers all combinations of contexts and during synthesis previously unseen combinations need to be dealt with. Therefore the parameters are shared between states by decision tree-based context clustering (Odell, 1995). Decision trees are constructed separately for excitation, spectrum, aperiodicity and duration.

2.3.3.4 Parameter Generation

At synthesis time an input sentence is converted into a sequence of context-dependent phonemes (see section 2.3.3.1). Speech parameters (excitation, spectral and duration) are then generated from the corresponding trained HMM-based models.

The core enabling technique for generating speech parameters from the HMM-based models is the ability to generate a perceptually smooth speech trajectory by taking into account constraints between static and dynamic properties of the trained statistical models (Tokuda et al., 2000). But in order to alleviate the problem that the generated speech parameter trajectory is too smooth, which makes it sound muffled, an extension to the generation framework that better takes into account the variation in the speech signal was developed in Toda and Tokuda (2007). The method in Toda and Tokuda (2007) use the global variance measure in section 2.3.3.2 to ensure that the generated speech parameter trajectory has variation across the utterance that is more similar to the variation in the natural speech.

2.3.3.5 Speaker and Style Adaptation

One important aspect of the HMM-based speech synthesis framework is the ability to adapt an existing synthetic voice to sound like a specific target speaker with only a small amount of target speaker data (Yamagishi et al., 2009). The adaptation together with the ability to share speech data between different speakers, often termed “average voice models”, removes the requirement for the phonetic coverage to be recorded from a single speaker (Yamagishi et al., 2007, 2009).

There exist several different adaptation techniques for HMM-based speech synthesis. The general adaptation techniques come from the neighbouring field of automatic

speech recognition. Yamagishi et al. (2009) conducted an empirical investigation of the performance of different adaptation algorithms for speech synthesis. Their results showed that the best adaptation performance was given by a constrained SMAPLR (CSMAPLR) combined with a posteriori MAP estimation. The algorithm adapts the mean and variance of the Gaussians in the clustered decision trees of the original voice to better match the target speaker (Yamagishi et al., 2009).

The average voice and adaptation techniques enable creating voices from non-conventional speech resources. Yamagishi et al. (2010) showed that they could create hundreds of different voices from speech data that was recorded with a variety of microphones and differing recording conditions. Not only the speaker identity can be adapted, but also the speaking style can be adapted. Tachibana et al. (2006) utilised the adaptation technique to adapt a voice with a neutral speaking style into voices with joyful and sad speaking styles.

2.3.3.6 Challenges for Conversational HMM Synthesis

The STRAIGHT (Kawahara et al., 1999) vocoder used in HTS is well capable of representing modal speech, but has limitations in representing breathy and creaky voice qualities, a problem addressed in Cabral et al. (2008); Silén et al. (2009). A better vocoder is a requirement for handling all aspects of conversational speech, e.g. laughter, but the STRAIGHT vocoder was hypothesised to preserve a sufficient degree of the phonetic properties of our conversational speech data to allow us to synthesise natural-sounding conversational characteristics.

Given that large amounts of accurately transcribed and phone aligned conversational speech data can be time consuming to obtain, the speaker and style adaptation techniques described in section 2.3.3.5 offered a potential short-cut to achieve conversational style synthetic speech from a limited amount of spontaneous speech data. However, in Lee et al. (2010) and in the pilot study in section 4.2.3 the result of adaptation showed that the listeners did not perceive a favourable distinction between the original read aloud voice and the adapted "spontaneous" voice. The reason behind these results is discussed in section 4.2.3.3, but the consequence was that we focused on speaker-dependent HMM-based speech synthesis.

In natural speech, listeners can hear the difference between when someone is speaking spontaneously or is reading aloud from a script (Blaauw, 1992, 1994; Laan, 1997). The limitations of the STRAIGHT vocoder and the negative findings of adaptation in Lee et al. (2010) and in section 4.2.3 suggested that the main challenge for conversa-

tional HMM-based speech synthesis is to preserve the phonetic properties that allow people to distinguish between natural spontaneous and read aloud speech.

The challenge was addressed by making a larger amount of conversational speech data available to allow building high quality voices from only conversational speech. The phonetic coverage of the conversational voice was boosted by blending conversational and read aloud speech data to further improve the quality of the voice. Then we made perceptual experiments to evaluate whether HMM-based voices trained from a substantial amount of conversational speech data could preserve a distinction between read aloud and conversational speech.

2.4 Conversational Speech Synthesis

Conventional synthetic voices have too limited expressive range to be useful for applications that require interacting in a more believable manner (Loyall, 1997; Traum et al., 2008; Romportl et al., 2010). To make synthetic voices suitable for believable characters, the voices need to be able to express a wider range of the speech phenomena found in human conversation. In sections 2.4.1 and 2.4.2 we will review previous approaches of making synthetic voices exhibit more conversational characteristics.

2.4.1 Synthesising Dialogue Acts

Of the work in conversational speech synthesis the approach in Campbell (2005, 2007) stands-out from other research by utilising a much larger corpus of conversational speech.

The speech corpus was recorded by letting volunteers carry a microphone and recording device with them during their everyday life and thereby capturing their everyday conversations. One woman was recorded over a period of five years, resulting in 600 hours of recordings (Campbell, 2007).

The time taken to record so much data is not feasible for the development of every new voice, but it provides an interesting dataset for learning how to utilise conversational speech for synthesis.

All the speech data was transcribed manually and split into utterances. The utterances were classified into two main types based on whether their contents were primarily propositional or affective. Half of the utterances were perceived as having primarily affective content that signalled speaker state (mood, emotions, health, involvement), or

speaker listener relationship (friend, stranger, formal, informal) (Campbell, 2005).

The affective utterances were further classified into dialogue acts: greeting, question, response, apology, backchannel, objection, suggestion, etc. The speaker state and speaker-listener relationship were represented in a simplified form as plus or minus "active and motivated" and plus or minus "friend/friendly" (Campbell, 2005).

Campbell (2005, 2007) argued that propositional content can be synthesised with the conventional speech synthesis methods, so they focused on synthesising the utterances with affective content.

Campbell (2005) argued that when synthesising greetings, backchannels, short confirmations etc., it is more important that the utterance has the prosodic properties of a greeting or backchannel to convey the appropriate pragmatic function, rather than just having the sequence of segments that form e.g. the word *"right"* or the phrase *"Hi, how are you doing?"*. The target sequence to synthesise was therefore not an orthographic word sequence. Instead, the utterance classifications were used as top-down targets to guide selection: the dialogue act (e.g. greeting), the speaker state, and who was speaking to whom (e.g. friends), limited the target phrases to a small set of phrases from which a token was selected. Campbell (2007) claimed that this phrase level selection made each isolated utterance sound natural.

In our opinion, keeping the lexical content underspecified and selecting whole phrases based on the affective content is a sensible idea for greetings, backchannels and short confirmations. But, despite five years of recorded data from one person the method was not sufficient to synthesise what she said in the sixth year (Campbell, 2007), which highlights the necessity of sub-word modelling of speech for synthesis.

2.4.2 Synthesising Filled Pauses and Hesitation

The integration of spontaneous or conversational characteristics into primarily propositional utterances has been addressed in a small number of approaches that will be reviewed in this section. All of them focused to some extent on filled pauses and the associated hesitation described in section 2.2.2.

The only approach described in this section that utilised spontaneous speech directly in the synthetic voices was Sundaram and Narayanan (2002). The other two approaches Cadic and Segalen (2008) and Adell et al. (2006, 2007b, 2010) based their models on analyses of spontaneous speech phenomena, but the speech used for synthesis was acted prompts.

Although Sundaram and Narayanan (2002) utilised spontaneous speech from a lecture, they only used fifty utterances to build a limited domain voice. They inserted filled pauses, breathing and laughter into utterances and showed that sentences with these phenomena were more likely to be confused with natural speech. They treated the filled pauses, breathing and laughter as tokens, but did not go into detail about how to select an appropriate token.

2.4.2.1 Hesitation 1

Cadic and Segalen (2008) designed sentences to cover word endings in French to model the transition from neutral speech into filled pauses and laughter for unit selection. They defined a speech sequence as consisting of neutral speech, followed by an anticipation phase, a paralinguistic element (filled pause or laughter), a return phase, and back to neutral speech. The anticipation phase was limited to consist of the ending of a word: the last consonant of a word followed by any other segments in the same word, motivated from a synthesis and not speech perspective in that consonants are better concatenation points than vowels.

From a corpus of text, Cadic and Segalen (2008) found that 200 word ending types covered more than ninety percent of their word ending tokens. The 200 word endings were included in sentences and a speaker was asked to read them twice: once with a filled pause, and once with laughter. The speaker was instructed to read aloud in a neutral manner up to the anticipation phase (the word ending) and then laugh or hesitate as naturally as possible. A unit selection voice was built and the anticipation phase and the hesitation or laughter was automatically concatenated into synthetic utterances. A perceptual evaluation showed that including an anticipation phase made the utterance sound more natural than just inserting the filled pause or laughter between silent pauses.

The approach in Cadic and Segalen (2008) shows the advantage of pre-selecting text to achieve coverage of conversational speech phenomena. But, in our opinion, the short stipulated anticipation phase seems to result in hesitation and laughter that are rather limited compared to the hesitation and laughter in natural conversation or in well acted speech.

2.4.2.2 Hesitation 2

In a series of papers Adell et al. (2006, 2007b, 2008, 2010) have addressed the modelling of filled pauses and associated hesitation for speech synthesis. They analysed duration and F0 patterns of filled pauses and surrounding context in spontaneous speech, and used the result to model F0 and duration targets for unit selection speech synthesis. We will not report their analysis results in detail, it suffices to say that their findings were consistent with the properties of filled pauses and associated hesitation in section 2.2.2: the filled pause itself was on average longer than other syllables, generally had a lower F0 than the rest of the utterance, and the syllable preceding the filled pause was prolonged (Adell et al., 2010).

A set of transcripts of spontaneous speech containing filled pauses were recorded by two voice talents. A comparison between these acted filled pauses with natural filled pauses showed some important similarities, but Adell et al. (2007b) also pointed out that one of the voice talents had less natural-sounding filled pauses and hesitation. In order to avoid coarticulation problems that arose with a small set of filled pauses (without a stipulated anticipation phase as in Cadic and Segalen (2008)) they were in synthesis always inserted between silent pauses (Adell et al., 2007b).

In Adell et al. (2010) a unit selection voice was built from 10h of read aloud speech, plus an additional 57 sentences containing filled pauses that were read aloud/acted by the same speaker. An evaluation showed that synthetic speech with and without filled pauses were perceived as equally natural.

The approach in Adell et al. (2010) is based on models of natural spontaneous speech. They evaluated the naturalness of their synthetic speech. However they did not attempt to evaluate to what extent the speech successfully synthesised a conversational style or conveyed a pragmatic function. Given that they used acted filled pauses and a large corpora of neutral speech this is a key weakness in this work.

2.4.2.3 Predicting Filled Pauses

The long term goal of a conversational speech synthesis system is to generate speech with appropriate content at an appropriate time in the conversation. This means that the content needs to predicted. Given a representation of the propositional content, e.g. a sentence, plausible placements of the wrappers, e.g. the filled pauses, can be generated.

In Sundaram and Narayanan (2003) filled pause insertions in text were modelled

by creating a list of word tokens and part-of-speech sequences that were frequently followed by a filled pause (mainly function words, such as: *a*, *and*, *but*, *the*). Then each part-of-speech and word token sequence was encoded in a finite state network. Given an input text with a matching part-of-speech or word sequence a filled pause was inserted. Breathing was inserted heuristically at the beginning of sentences, between phrases and before half of the *um*:s. No evaluation of the predicted insertions was made, but an example of transformed input from Sundaram and Narayanan (2003, p. 4) is shown below:

INPUT: "Might as well talk about it right now"

Transformed-INPUT: "[BREATHE IN] Might as well talk about it [UM] right now."

In a similar experiment, Adell et al. (2007a) modelled filled pause insertions in text with decision trees. The features used to build the decision tree were: current word, bigram probabilities of word pairs, and part-of-speech of previous, current and next word. The set of words was limited to only forty candidate words, motivated by that the ten most frequent words were followed by over 50% of the filled pauses in their multi-speaker corpus of spontaneous speech. An evaluation of their filled pause insertions on test data showed 97% precision and 58% recall when using the above features.

The use of a limited set of function words seemed sufficient to generate plausible filled pause placements in text. In Andersson et al. (2010a) we developed methods for insertions of both filled pauses and discourse markers, and in section 5.5 we will describe an evaluation of how the predicted insertions affected the perceived quality of the synthetic speech.

2.4.2.4 Conversational Speech Synthesis Summary

The general aim of the work in conversational speech synthesis is to extend the limited behaviour of conventional synthetic voices and synthesise a richer variety of the speech phenomena found in human conversations. Our approach outlined in sections 2.2 and 2.3 of augmenting the conventional read aloud voices with speech from a spontaneous conversation lies roughly inbetween the previous approaches to conversational speech synthesis that are described in sections 2.4.1 and 2.4.2.

In sections 2.4.1 and 2.4.2 we described two very different approaches to conversational speech synthesis. In section 2.4.1 we described how Campbell (2007) argued

that utterances with primarily affective content, e.g. greetings or backchannels, should be selected as whole utterances based on their affective content (e.g. informal or formal greeting) rather than their phoneme sequence. The selections of utterances were made from a 600h corpus of one person's spontaneous conversations. In section 2.4.2 we described how Adell et al. (2010) and Cadic and Segalen (2008) added filled pauses to their synthetic voices by creating models for selecting filled pauses from recordings of acted filled pauses. Adell et al. (2010) showed that they could synthesise utterances containing filled pauses that sounded as natural as synthetic utterances with only propositional content.

2.5 Conclusion

The main challenge for conversational speech synthesis is to enable the synthetic voices to synthesise a wide range of conversational characteristics while maintaining the quality that can be achieved with conventional "neutral" synthetic voices.

We will investigate to what extent we can utilise speech from a spontaneous conversation to synthesise natural-sounding conversational style speech with the unit selection and HMM-based speech synthesis methods. We will focus on two types of speech phenomena that have generally been neglected in conventional speech synthesis: the discourse markers (e.g. *yeah*, *you know* or *'cause*) and filled pauses (*um* and *uh*) that were described in sections 2.2.1 and 2.2.2, because by synthesising speech where discourse markers and filled pauses are wrapped around propositional content, the synthetic voices can express both affective and propositional information, e.g. certainty or uncertainty about a topic, in the same way humans express it in spontaneous conversation.

In chapter 3 we will describe the recording and analysis of the spontaneous conversation that was used to build the synthetic voices in chapter 4. Chapter 4 describes the building of the synthetic voices and outline the details of our blending techniques. Chapter 5 describes the perceptual evaluations of the voices.

Chapter 3

The Speech Data

The first part of this chapter will describe how the conversations used in this work were recorded and transcribed, and how a subset of the conversations was selected for use in speech synthesis. The second part of this chapter will describe the linguistic and phonetic properties of the selected subset of speech from a conversation, in comparison to a more conventional speech synthesis resource of carefully read aloud sentences.

3.1 Independent Contribution of the Author: Eliciting and Processing Conversational Speech

Part of the recording, processing and analysis of the speech described in this chapter has been used for the joint publications in Andersson et al. (2010a), Andersson et al. (2010b) and Andersson et al. (2012). This section outlines the current author's independent contribution to the speech data collection and analysis. All the methodological decisions regarding speech data collection, preparation and analysis were made by the current author.

- The eliciting of conversational speech from the three voice talents in section 3.2 was made by the current author.

- The transcription of the conversations in section 3.4 was carried out by the current author.

- The modification of the segmentation in section 3.4.5 and the analysis of the speech data in section 3.5 were carried out by the current author.

The majority of the acoustic and linguistic analysis was carried out using the tools available in the CereVoice (Aylett and Pidcock, 2007) and HTS (Zen et al., 2007) speech synthesis systems. The remaining analysis was carried out using the available signal processing software installed at the Centre for Speech Technology Research, at the University of Edinburgh.

The majority of the figures and tables in this chapter were not part of the joint publications in (Andersson et al., 2010a,b, 2012). The figures and tables that have been part of our joint publications are generally presented in this chapter in modified and more detailed versions.

3.2 Recording Spontaneous Conversation

In total we recorded three voice talents:

- Heather, a Scottish female in her early twenties
- Roger, an English male in his forties
- Johnny, an American male in his late thirties

These three voice talents were originally cast for speech synthesis projects unrelated to this thesis. Heather was cast by CereProc, Roger was cast by The Centre for Speech Technology Research (CSTR) and Johnny was cast by David Traum's group at the USC Institute for Creative Technologies. To ensure matching recording conditions to the voice talents' previous recordings of read aloud sentences, the author was assisted by the previous recording technicians Chris Pidcock (Johnny and Heather), Yolanda Vazques-Alvarez (Roger) and Ziggy Campbell (Roger). The technicians setup the recording tools and the author managed the recordings during the sessions with the voice talents.

The conversations with Heather and Roger were used to pilot the general approach of utilising conversational speech for synthesis. The recorded conversations with Heather and Roger therefore only lasted approximately an hour each, which gave about 20min of speech from each voice talent. This speech data was used in the pilot speech synthesis experiments in sections 4.2.3 and 4.3.2. The results from these pilot experiments motivated the recording of the longer, 7h, conversation with Johnny described in section 3.2.2.

The recordings were made in separate sessions for each voice talent and the author of this thesis was the conversation partner in all the recordings. The recordings were made in a recording studio where the voice talent was positioned inside a booth, but had eye-contact with the author through a window. The voice talent and the author spoke to each other via microphones and headphones. The speech of the author and of the voice talents were recorded on separate channels. All the voice talents had been recorded before, reading aloud sentences for synthesis. We used the same studios and microphones when recording the conversations to facilitate comparisons between, and blending of, spontaneous and read aloud speech. The conversations were recorded with 48KHz sampling rate and 16bit sample depth.

Eliciting spontaneous conversation from a paid voice talent has advantages and disadvantages. The disadvantage is the ecological validity of the artificial situation. But, in our experience, it is not difficult to get people to talk about themselves and their interests in a friendly environment. The advantages are the controlled recording environment and that the voice talents could be requested to a) not "put on" different voices to portray another person, such as their partners or children, and b) not talk only about themselves, but also ask about the author's life and interests. Although such explicit requests are artificially imposed constraints, the impact on spontaneity is minor. Below we show an example where Johnny adhered to such a request, to give some impression of how the requests affected the speech:

- *I'm not gonna do a damn voice but damn it if I don't want to*
  (When he felt an urge to mimic an old girlfriend.)

In sections 3.2.1 and 3.2.2 we give an overview of the conversations with the three voice talents. The examples from the conversations given in sections 3.2.1 and 3.2.2, were not all used in the synthetic voices described in chapter 4. Sections 3.4.1, 3.4.2, 3.4.3 and 3.4.4 outline how speech from the conversations was transcribed and selected for use in the synthetic voices.

3.2.1 Speaking with Heather and Roger

The recorded conversation sessions for Heather and Roger lasted approximately one hour each. Two examples from the conversations are shown below (the first is from Heather and the second is from Roger):

- *although I was really lucky my [pause] my supervisor was great [pause] the only*

[pause] the only thing I could say against her was the fact she's a Hibs supporter which uh [pause] definitely counts against her

- *were you uh [pause] serious when you were suggesting continuing the conversation or was that a subtle ploy to get me back into the uh [pause] into recording studio*

Sections 3.3 and 3.4 outline our method of transcribing the conversations. The conversations gave 392 transcribed utterances from Heather and 265 utterances from Roger. The conclusion drawn from the pilot recording sessions with Heather and Roger was: studio recorded spontaneous conversation is a straightforward method to obtain conversational speech for speech synthesis. These conversational utterances were used for the pilot voice building experiments described in sections 4.2.3 and 4.3.2.

Although our approach was to utilise blending of conversational and read aloud speech to address the lack of phonetic coverage in spontaneous speech resources, this approach still requires a sufficient amount of conversational speech as outlined in sections 2.3.2.2 and 2.3.3.6. Therefore, we recorded the longer conversation with Johnny. Figure 3.4 shows that better coverage can be achieved in spontaneous speech resources up to about one thousand utterances for diphones, and at least two thousand for quinphones.

3.2.2 Speaking with Johnny

The conversation with Johnny[1] was recorded in three sessions spread over a period of five days and lasted a total of approximately seven hours. The speech from the conversation with Johnny was used in the final unit selection and HMM-based synthetic voices described in chapter 4.

The conversation with Johnny mainly focused around the voice talent's professional career as an actor and director, former boxing career, his family and life in general in the U.S. Below we show a short sample from the conversation, where the author and Johnny discussed filled pauses in acted and spontaneous speech. In the example, spontaneous filled pauses (*um* or *uh*) and meta-communication about filled pauses (quoted and bold faced) are intermingled (only Johnny's speech is shown):

- *yeah [pause] yeah [pause] no and that's cool I mean that's the thing that that's [pause] that's weird too because even like in the script [pause] um [pause]* ***"um"***

[1]The recording of Johnny's speech was made by the author while visiting The USC Institute for Creative Technologies (http://ict.usc.edu).

is usually not [pause] scripted [...] *I'll [pause] I'll put it in there as part of [pause] my speech [pause] but it's not necessarily in the script*

- *um but I've never really had it [pause] scripted [pause] because it's uh [pause] an **"um"** is almost something you have to earn [pause] you know what I mean [pause] it's like uh [pause] like uh [pause] Harold Pinter was like um [pause] uh [pause] he's a playwright and [pause] talked about how you had to earn your pauses [pause] and so he would [pause] specifically put pauses in there but you can't just pause outta nowhere*

The explicit request to the voice talent to ask about the authors life and interests often resulted in a few "uninteresting" questions (e.g. *how tall are you*), before leading back to a less staged interaction (only Johnny's speech is shown):

- *yeah do you guys [pause] use like for weight do you guys talk about stone*
- *kilograms [pause] do you know what a stone is*
- *[...] okay it just confused the hell outta me and of course I couldn't just [breathe in] spend [pause] a minute and a half on the Internet and get that figured out*

Even without explicit requests the interpersonal exchange of conversations affects the speech of the people involved. Johnny occasionally used the expression *the whole nine yards* until the author, unfamiliar with the expression, asked him about it:

- *yeah [pause] but the whole nine yards I don't even like that I say that [pause] and it's just now being brought to my attention by you [pause] so*
  (After that, Johnny did not say *the whole nine yards* again.)
- The author used the British expression *knackered*, which was unfamiliar to Johnny, who included it in his vocabulary: *yeah [pause] I love that word knackered by the way [...] yeah [pause] I've never heard it before [...] so it's been like every other word at the house*

Hence, the sessions recorded with Johnny gave a substantial amount of spontaneous conversational speech to use for speech synthesis. The processing and analysis of this speech data is described in detail in sections 3.3, 3.4 and 3.5. In total we obtained 2120 conversational utterances containing 75min of phonetic material. Table 3.2 summarises the contents of the speech data. This data was used to build the voices for the perceptual evaluations in chapter 5.

3.3 Representing Meaning of Conversational Speech

We used orthographic transcription to represent the meaning in conversational speech. The general hypothesis was that an orthographic transcript of a conversation would provide a speech resource for building synthetic voices with natural-sounding conversational characteristics. The main motivations behind this hypothesis were:

- The orthographic word (sequence) together with utterance position was hypothesised to capture the prototypical meaning and associated phonetic properties of discourse markers and filled pauses. For example, *yeah* as a stand-alone backchannel, in the confirmation *yeah yeah yeah*, or in the beginning of a longer turn *yeah I feel kind of dirty afterwards*.

- An orthographic transcript also implicitly captures the more expressive nature of many other words in conversational speech. The following samples illustrate how orthography in the typical case capture the speaker's positive or negative opinion about a topic:

  - *I [pause] fucking hate commercials [pause] I can not stand it [pause] oh drives me insane*
  - *for me I really love to do what I do [pause] I love it*
  - *I don't think that [pause] celebrity and politics [pause] need to be related*

The orthographic transcripts of speech from a conversation therefore enabled us to focus on the integration of a wide variety of discourse markers and filled pauses together with propositional content. This allowed us to synthesise speech capable of expressing e.g. certainty or uncertainty about a topic in a natural manner and thereby give our voices a more conversational style than conventional synthetic voices. As in the examples taken from our recorded conversation where the propositional content is in bold face and discourse markers and filled pauses are in italics:

- *oh yeah* **it's great exercise** *so*
- *yeah* **I can see that**
- **whether successful or not I I aim for that** [pause] *you know*
- *and you know* **it's just like any other job you hire people you like working with**

- *um* [pause] *no I uh* [pause] *uh* **I moved up for acting**
- *exactly because* **he's he's Robert de Niro**
- *wow* **that's really cheap**
- *well* **she doesn't know about that one**
- *yeah* [pause] *so* [pause] *and* **it's not high end high quality food** [pause] *I mean* **it's** [pause] **beans and cheese and ground beef and tortillas** *and yadda yadda*

3.4 Transcription, Selection and Segmentation

In this section we will describe how the speech was transcribed and give examples of speech sequences that were selected or rejected for use in the synthetic voices built in chapter 4. The speech from the conversations was manually transcribed. This method was selected over using automatic speech recognition to eliminate erroneous transcripts as the predominant error source in the synthetic voices.

The decision of which utterances to use to enable natural and conversational synthetic voices was based on:

- concatenation errors in pilot unit selection voices
- results from forced alignment
- listed pronunciations in the lexicon.

Only utterances that were considered for use in speech synthesis were transcribed. Utterances that contained word fragments, mispronunciations, heavily reduced pronunciations or mumbling were therefore not transcribed and not used in the synthetic voices. When possible the utterance boundaries were placed so as to exclude any immediately preceding or succeeding laughter, sigh, throat clearing, etc.

The selection of utterances was less strict for Heather and Roger than for Johnny, because Heather and Roger were used to pilot the general approach of utilising speech from spontaneous conversations for speech synthesis. These less strict selections provided valuable insight into how heavily reduced pronunciations and laughing speech (where the voice talent laughed and spoke at the same time) affected both the forced alignment and subsequent synthetic voices. The utterances where Johnny "put on" different voices to mimic a third person, e.g. his wife, children or friends, were excluded.

This was done to retain speech that represented the voice talent's "normal" speech, speech that could be considered his consistent spontaneous conversational speaking style.

3.4.1 Splitting into Utterances

The only utterance level analysis made by the CereVoice system was between intermediate and utterance final phrase boundaries. The utterance final boundary was always assigned the same phrase boundary category. The intermediate boundary was assigned when an utterance internal silent pause between 50-250ms was automatically detected. If an utterance internal silence was longer than 250ms, a phrase final boundary was assigned. Therefore, the conversation was not split at every silence, but instead we aimed at splitting the conversation at the end of a statement, question etc. But, no annotation of dialogue act was made.

For an isolated read aloud sentence it is very easy to determine the beginning and end, because they exist. In a conversation people do not speak in isolated sentences and the notion of utterance beginning and end is more complicated. The splitting of a conversation into utterances was not always difficult, many times it was a straightforward task, in e.g. stand alone and (fairly) grammatical questions, statements, short responses, confirmations, and most backchannels. But sometimes it was more problematic:

- Some discourse markers (e.g. *and*, *and uh*, *and so*, *so*) were used to signal that the speaker did not consider the current topic to be closed, which could result in long sequences of speech without silent pauses in connection with a clear phrase final boundary. As a general guideline we attempted to keep the utterances below forty words, and split these longer sequences at the best available silent pause, based on language content and acoustics. The example in figure 3.1 shows such a long sequence of speech and how it was split.

- In a conversation the participants sometimes interrupt each other. Clear interruptions that resulted in word fragments or clearly unfinished utterances were excluded. But discourse markers and filled pauses also offered possibility for speaker and topic changes, based on the fact that many speaker and/or topic changes occurred after *and*, *and uh*, *um*, *uh* even though the current topic was not necessarily closed off.

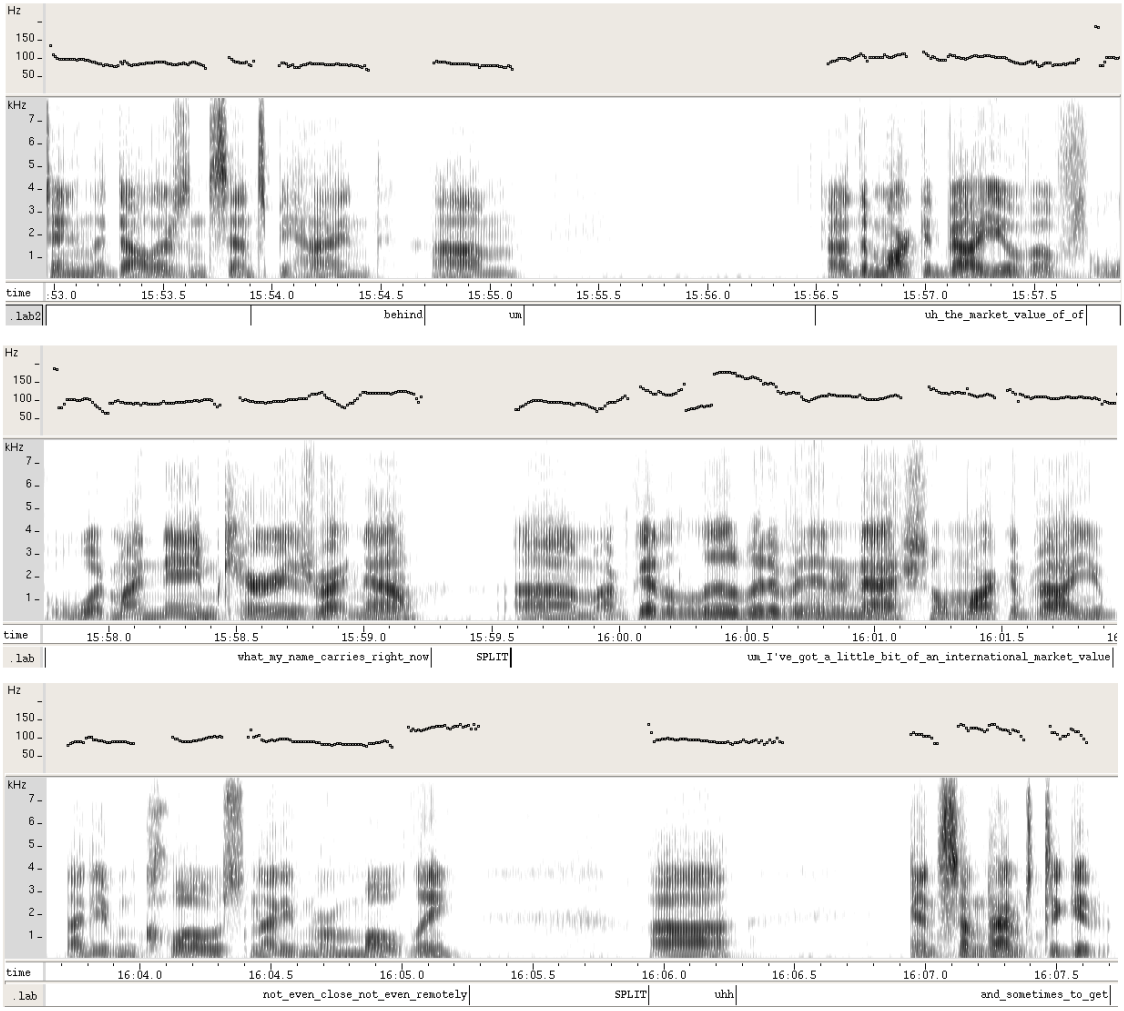

Figure 3.1: The figure shows an example of splitting a long speech sequence into more manageable utterances, based on acoustics and language content. The top pane shows a decision to not split, and the two bottom panes show where splits were made. The speech sequences shown in the panes are bold faced in the transcript of the whole sequence below: *no it's not even about doing it better um what it is for me is understanding the dollars* ***and cents behind [pause] um [pause] uh the market value of of what my name carries right now*** **SPLIT** ***um I've got a little bit of an international market value*** *but I'm not a big star at all [pause]* ***not even close not even remotely [breathe in]*** **SPLIT** ***uh [pause] and sometimes to get*** *films funded properly [...]*

3.4.2 Pronunciation and Enunciation

The use of forced alignment required a close match between the stipulated phonemic pronunciations available in our pronunciation lexicons and the actual pronunciations in our recorded conversations. Adding words, changing word order or omitting words in the transcript with respect to what was actually said is devastating for forced alignment. But transcribing speech is a very different task from listening to or engaging in a conversation, and it is easy to interpret what was being said and make utterances more grammatical than they actually were, by e.g. changing word order, omitting function word repetitions or filled pauses (Lickley, 1995). This required careful attention to the detail of the utterances, which made the total time for transcribing and selecting speech from the seven hour conversation with Johnny take approximately 1-2 months of fulltime work.

In general the most problematic words to transcribe for synthesis were the function words (e.g. did they say *a* or *the* or nothing), and the discourse markers; both one word e.g. *'cause*, *probably* or *especially*, and longer ones such as: *you know what I mean* or *at the end of the day*. The expressions *you know what I mean* and *at the end of the day* were frequently used by Johnny, but it was often not clear, either from listening or from the spectrogram, which of the words or phonemes were there or not (but the "gist" of the expression was clearly there). The expressions *you know what I mean* and *at the end of the day* were only selected for use in synthesis when all the words were clearly present. But, heavily reduced function words were included in the selected utterances, because of their very frequency they must be included to retain any spontaneous data at all.

In an analysis by (Johnson, 2004) of heavily reduced pronunciations in American English conversations, they listed several of the pronunciation variants of e.g. *because* and *probably* that were encountered in our spontaneous speech data. Whereas *because* was often pronounced close to the citation form, *'cause* was sometimes heavily reduced and only pronounced, as in an example from Johnson (2004), as [k^hz]. In figure 3.2, examples of Heather's reduced *'cause* are contrasted with her pronunciation of *because*. Figure 3.2 also shows the result from the phonemic forced alignment where the phoneme identity, or quality of the vowel, is at best questionable, but the word boundaries are correct, which actually would make them usable as word representations but not, as did happen in the pilot voices, as phonemic units in (m)any other word contexts. Such heavily reduced pronunciations that deviated substantially from the listed

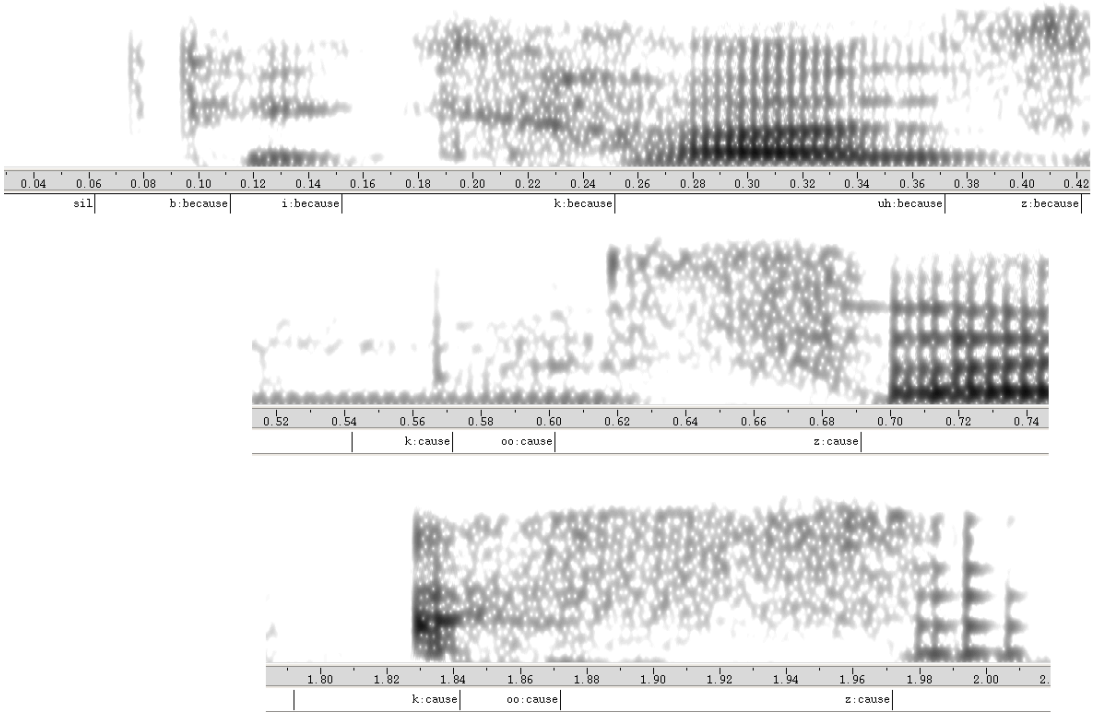

Figure 3.2: Examples of forced alignment and actual pronunciations of *because* (top pane) and reduced variants of *'cause* pronounced as roughly [k^{h}z]. This [k^{h}z] was not listed as valid pronunciation variant in our pronunciation lexicon. As stated in Clark et al. (2007), forced alignment is often more consistent than accurate in aligning phone boundaries, this is visible in some of the displayed boundaries in this figure, also for the fully pronounced *because* (bikuhz) in the top pane.

pronunciations in our lexicon were therefore excluded in the more restrictive selection of Johnny's speech. Less reduced variants of *because* and *especially* were transcribed as *'cause* or *specially*, but pronunciations of *probably* as [prɑli] or *'cause* as [k^{h}z] were excluded.

Only words that were missing from the pronunciation lexicon, mainly proper names, were added to it. No alterations to the existing lexical entries were made.

3.4.3 Transcribing Filled Pauses and Other Non-lexical Items

Filled pauses have been converted to orthographic notation in a few different ways by different authors, but there was no convincing evidence that motivated transcriptions such as *um*, *u:m*, *uh*, *u:h* (as in Clark and Fox Tree, 2002) or *um*, *umm*, *uumm*, *uh*, *uhhh*, *uuuh* (as in Ward, 2006), and hence they were transcribed in this thesis as

"just" *um* or *uh*. In the CereVoice lexicon the filled pauses were represented as a vowel only for *uh* and a vowel followed by /m/ for *um*. The vowel quality of the filled pauses differed slightly between the lexicons for Scottish, English and American accents. Whether that vowel quality corresponded to the filled pauses in our speech was not further elaborated on, because most filled pauses seemed at least to have the same vowel quality, and phonemic representations of non-lexical words mainly act as "place holders" for forced alignment and speech synthesis. What matters is in which contexts these phonemes/place-holders occur. Figures 4.5 and 4.4 show that this place-holder method preserved both duration and vowel quality of the filled pauses in the synthetic speech.

The same place-holder guideline was applied to other items that lacked a "correct spelling" but had phonemic properties that were different from the same phonemes in other word types, e.g. the backchannels *uh-huh* and *mhm* and the conversational "grunts" (e.g. *hmm*, *huh*). For example, *mhm* was represented in the lexicon as /mhəm/, but those phonemes should, in unit selection, only be used to synthesise *mhm*.

Laughter is an integral part of conversations, and the boundary of what is laughter and what is speech is not always clear. Stand-alone laughter was always excluded from the selected utterances, but for Roger and Heather laughing speech (speaking and laughing at the same time) were included. An example of laughing speech in the word *Glasgow* is shown in figure 3.3. In the pilot unit selection voice (in section 4.3.2) this example of laughing speech did have a positive impact on the resulting synthetic utterance: *um [pause] but yeah I think I prefer Glasgow*, but this was more to do with "limited domain" factors than sub-word unit selection. Therefore, laughing speech was not selected for Johnny's synthetic voices.

3.4.4 Speech Disfluencies

Speech disfluencies are very frequent in spontaneous speech and were included in the selected utterances, except when they contained word fragments or mis-pronunciations (e.g. pronouncing *ball* as *pall*). A few transcribed examples (from Johnny) of disfluencies (bold faced in the examples) that were included in the final synthetic voices are:

- *yeah* ***it's it's*** *a significant amount of swelling [pause] um [pause] more than like I'd say a bruise*
- *but um [pause] she's not even good* ***at [pause] at [pause] at*** *hiding ulterior*

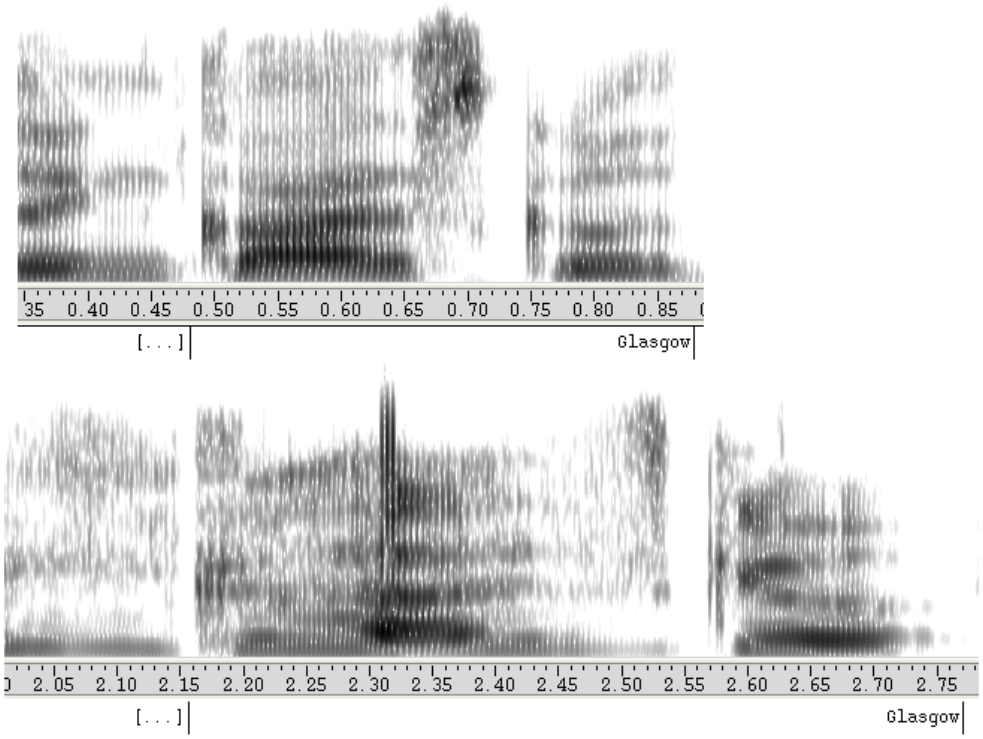

Figure 3.3: Spectrograms of *Glasgow* in modal (top pane) and laughing speech (bottom pane)

motives you know what I mean she's very blatant about it

- *and I think that's really [pause]* ***that that that that*** *conflict within yourself [breathe in] versus like you see in the movies all the time*
- *yeah no* ***I'm*** *I go like hey where's the chips at*
- *and then you can go [pause] oh this is actually their job* ***and this is*** *and they want tips [pause] so [pause] this would actually help them out by letting them carry my luggage up you know*

3.4.5 Segmenting the Conversational Speech

In the previous sections in this chapter we described how the speech from the spontaneous conversations was transcribed and selected to:

- provide speech that represented the speakers' consistent spontaneous speaking

style

- provide speech that, to a large extent, could be automatically force aligned.

The CereVoice implementation of forced alignment described in section 2.3.1.1 was used. Forced alignment using only spontaneous speech did not provide sufficiently accurate phoneme alignment for speech synthesis for any of the voice talents, despite the more restrictive utterance selection and substantially more speech from Johnny. To improve the alignment of the spontaneous speech, without negatively affecting the alignment of the read aloud speech, we initialised the HMM-models for the spontaneous speech from trained read aloud models from the same speaker as follows:

1. do forced alignment of the read aloud speech (as outlined in section 2.3.1.1)

2. slow down the spontaneous speech by 10% with SoundStretch (Parviainen, 2012), to better match the speaking rate of the read aloud speech (see section 3.5.2.3)

3. initialise the HMM-models for the spontaneous speech from semi-trained[2] read aloud HMM-models, including silence and short pause models

4. continue updating the HMM-model parameters with a further three iterations of Baum-Welch training using only the spontaneous speech

5. force align with the Viterbi algorithm as described in Young et al. (2006).

The method was developed from the speech of Heather and Roger for which it made an improvement to synthetic speech quality.

Apart from the improvement obtained by utilising the read aloud models for alignment, an important contributing factor to the generally good segmentation quality of the spontaneous speech was CereVoice's use of pronunciation variants with a better match to actual pronunciations than the citation form pronunciations, e.g. *and* as /ən/ or *around* as *aroun*.

[2] "Semi-trained" consisted of six iterations of Baum-Welch training. The complete forced alignment training in CereVoice do more training iterations, therefore we use the term "semi-trained" instead of "trained". This decision was largely made from a practical perspective and the alternative of intialising with fully trained models was not tested.

3.4.5.1 Segmentation Result

Table 3.1 shows the result of an evaluation of the forced alignment on ten randomly selected utterances from Johnny's conversational and read aloud speech. The automatically aligned phoneme boundaries were compared to manually corrected boundaries. Only boundaries that were considered erroneous by more than 25ms were adjusted. The total alignment error was about three times higher for the conversational speech. However, two misaligned /t/ segments in one conversational utterance accounted for 1500 ms out of the total 2085 ms error. The evaluation confirmed our impression that the forced alignment of the conversational speech was in general accurate, albeit not as accurate as for the read aloud speech, and that there were more gross alignment errors in the conversational utterances.

3.5 Comparing Read Aloud and Conversational Speech

The blending approach to conversational speech synthesis in this thesis utilised both conversational and read aloud data. In sections 3.5.1 and 3.5.2 we will compare our conversational and read aloud data, to show that blending them is possible. This blending will address the lack of phonetic coverage in the conversational data, while preserving the spontaneous quality of distinguishing speech phenomena in our conversational speech data.

In addition to the transcribed conversational utterances we had recordings of neutrally read aloud sentences available for all voice talents. The read aloud sentences were recorded by Strom et al. (2006, 2007) for Roger, and by CereProc for Heather and Johnny. These sentences were recorded to provide phonetic coverage of diphones in different segmental and prosodic contexts. The sentences were recorded in the same studios and with the same microphones as the conversations, and in the case of Johnny also around the same time as recording the conversation.

In the following sections we will quantify some of the linguistic and phonetic properties of the conversational and read aloud speech. The conversations from Heather and Roger only gave 392 and 265 utterances respectively. The conversations from Johnny gave a more substantial 2120 utterances. Therefore, the linguistic and phonetic analyses presented in sections 3.5.1 and 3.5.2 were made on Johnny's speech, but some references will be made to the speech of Heather and Roger. Table 3.2 gives an overview of the composition of Johnny's conversational and read aloud data. Part

utt	no. of phonemes	error (phonemes)	error (ms)	max error (ms)
RD_1	36	4	130ms	50ms
RD_2	55	3	175ms	100ms
RD_3	9	0	0	0
RD_4	19	2	75ms	50ms
RD_5	17	2	140ms	80ms
RD_6	32	0	0	0
RD_7	24	3	75ms	25ms
RD_8	36	2	50ms	25ms
RD_9	41	2	65ms	40ms
RD_10	44	0	0	0
Total	313	18	710ms	-
utt_id	no. of phonemes	error (phonemes)	error (ms)	max error (ms)
SP_1	78	3	85ms	30ms
SP_2	5	1	40ms	40ms
SP_3	35	4	155ms	70ms
SP_4	8	0	0	0
SP_5	143	12	1725ms	800ms
SP_6	3	0	0	0
SP_7	8	0	0	0
SP_8	11	1	50ms	50ms
SP_9	14	0	0	0
SP_10	42	1	30ms	30ms
Total	347	22	2085ms	-

Table 3.1: Forced alignment errors in Johnny's read aloud (RD) and conversational (SP) speech.

	Conversation	Read Aloud
utterances	2120	2717
word tokens	19841	22363
word types	2200	5026
syllable tokens	24657	30902
phone tokens	58332	75856
diphone types	1769	2483
quinphone types	37654	58867
total duration (incl. silence)	89min	106min
total duration (excl. silence)	75min	103min

Table 3.2: Overview of Johnny's conversational and read aloud data. The duration shows the amount of phonetic material, including or excluding utterance internal silent pauses. The diphone types include silences and lexical stress on vowels. The quinphone types include silences, but not lexical stress.

of the analyses in Sections 3.5.1 and 3.5.2 were published in Andersson et al. (2012) where the first author conducted all the linguistic and phonetic analyses.

3.5.1 Language Composition and Phonetic Coverage

The linguistic analysis in the CereVoice system provided an analysis mainly based on linguistic features extracted from the text of an utterance, such as phoneme identity, neighbouring phonemes, lexical stress and phrase position. The use of these automatically predicted features means that there is no need for manual mark-up and the features can be predicted also for the unseen text that we need to synthesise.

Table 3.2 shows that there was more read aloud than conversational data. In addition to this overall difference, the two datasets have differences in language composition and phonetic coverage that have consequences for our aim of integrating discourse markers and filled pauses with propositional content in synthetic utterances.

Figure 3.4 shows how phonetic coverage of diphone and quinphone types increases as a function of number of utterances in the read aloud and conversational data, where the benefit of pre-selecting sentences to achieve phonetic coverage of, in particular, diphones is illustrated. However, the read aloud utterances did not have better coverage of everything. Table 3.3 shows the twenty most frequent words in Johnny's conversational and read aloud data. Short function words, such as *the*, *a*, *of* or *to* were frequent

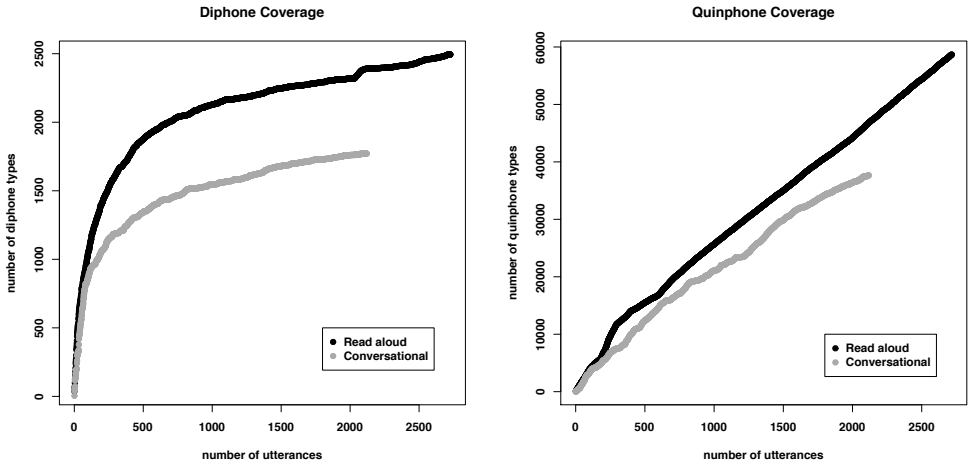

Figure 3.4: Diphone and quinphone coverage in the read aloud and conversational speech. As in table 3.2, the diphones include three levels of lexical stress on the vowels, making the phoneset contain 72 phonemes. If we exclude the lexical stress, the read aloud data contains 86% of the theoretically possible diphones, and the conversational data contains 70%. The quinphone coverage does not include lexical stress on the vowels.

in both datasets. The most frequent word in the conversational data was *yeah*, which occurred a mere three times in the read aloud data, and many other words, e.g. *know* and *so*, showed similarly large distributional differences between the read aloud and conversational data.

The reason for these distributional differences is that many of the frequent words in the conversational data are frequent because they were used to regulate the conversational flow, through discourse markers and backchannels, or express non-propositional content such as agreement or hesitation. Approximately thirty percent of Johnny's conversational utterances consisted of a single isolated word (e.g. 339 *yeah*, 167 *right*, and 54 *okay*) of which the majority were backchannels. The discourse markers and filled pauses were however mainly integrated with propositional content in longer utterances, and as table 3.4 shows, often occurred in the vicinity of the phrase or utterance boundaries. This distribution of discourse markers and filled pauses around phrase and utterance boundaries represents our speaker's means of structuring his speech in conversation to start, end or keep a turn.

	Conversational		Read Aloud	
rank	type	count	type	count
1	**yeah**	818	a	762
2	I	787	the	709
3	and	690	I	390
4	you	570	to	390
5	the	488	of	340
6	a	448	is	304
7	**that**	366	and	290
8	**know**	344	you	251
9	to	336	in	220
10	**uh**	318	**he**	204
11	**so**	302	it	193
12	**um**	292	**one**	192
13	it	291	with	167
14	of	278	**two**	165
15	**it's**	262	**we**	155
16	**but**	248	was	151
17	**like**	217	**three**	138
18	**right**	210	**on**	134
19	was	207	are	131
20	is	195	**they**	130

Table 3.3: The twenty most frequent words in Johnny's conversational and read aloud data. Non-overlapping words between the two columns are bold faced.

3.5.2 Phonetic Properties

This section will show overall acoustic phonetic properties of the read aloud and conversational speech data in table 3.2.

3.5.2.1 Energy

For the HMM-based synthetic voices in section 4.2 the spectral parameters were extracted as 39th order STRAIGHT (Kawahara et al., 1999) mel-cepstral coefficients. The 0th coefficient is a measure of the energy in the speech frame. Figure 3.5 shows

frequency	trigram	frequency	trigram	frequency	trigram
339	sil_yeah_sil	19	sil_and_I	12	um_sp_you
167	sil_right_sil	18	I_mean_sil	11	a_bunch_of
124	sil_yeah_sp	18	yeah_yeah_sil	11	and_uh_sil
118	sp_you_know	17	you_know_I	11	and_you_know
68	sil_um_sp	17	but_uh_sp	11	sil_oh_yeah
68	sil_you_know	16	sp_and_uh	11	that_sp_I
54	sil_okay_sil	16	and_uh_sp	11	what_it_is
53	yeah_sp_yeah	16	sp_and_then	11	yeah_I_mean
46	you_know_sp	16	sp_yeah_yeah	11	yeah_sp_no
43	you_know_what	15	sp_and_I	11	sil_no_sil
38	know_what_I	14	I_don't_know	10	sp_I_was
38	you_know_sil	14	it's_uh_sp	10	and_I_think
37	a_lot_of	14	sp_and_sp	10	sil_and_sp
37	sp_um_sp	14	when_I_was	10	sil_and_uh
37	sp_yeah_sil	13	sil_but_uh	10	sil_right_sp
36	sp_um_sil	13	sil_yeah_yeah	10	sp_but_uh
36	what_I_mean	13	uh_sp_I	10	sp_exactly_sil
27	sil_yeah_I	13	um_sp_I	10	sil_exactly_sil
27	sp_so_sil	13	um_sp_and	10	sil_nice_sil
23	sp_uh_sp	13	yeah_sp_exactly	10	sp_I_was
23	sp_yeah_sp	12	sil_and_so	10	sp_it_was
20	you_know_and	12	sil_I_mean	10	yeah_sp_I
20	sil_and_then	12	sil_yeah_no	10	yeah_yeah_sp
19	uh_sp_yeah	12	um_sp_but	10	yeah_yeah_yeah

Table 3.4: Trigrams occurring ten times or more in Johnny's conversational data. The trigrams include utterance beginning/end as "sil", and utterance internal short pauses as "sp".

the overall distribution of energy in the read aloud and conversational speech. Figure 3.6 shows the distribution of energy in the centre of the vowels in the read aloud and conversational speech.

The higher proportion of utterance internal silence in the conversational data reported in table 3.2 is visible also in figure 3.5 as a plateau in the lower energy region.

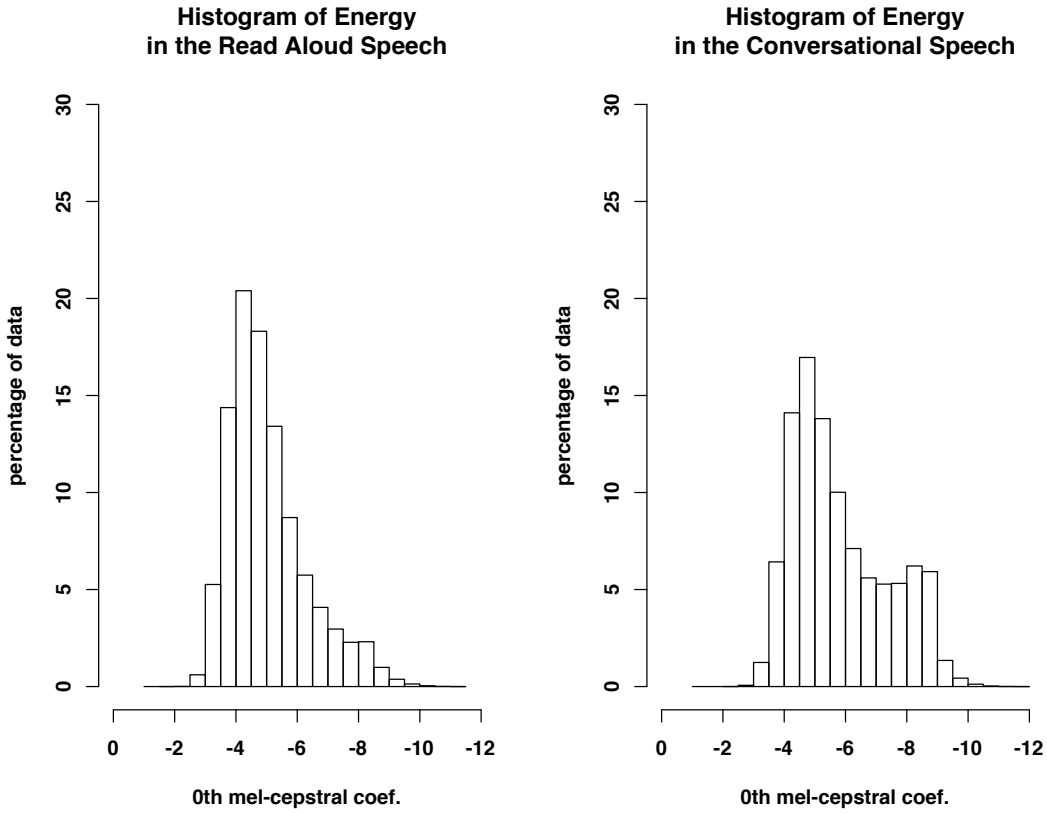

Figure 3.5: Histogram of overall energy distribution, measured as the 0th mel-cepstrum coefficient, in the read aloud and conversational speech.

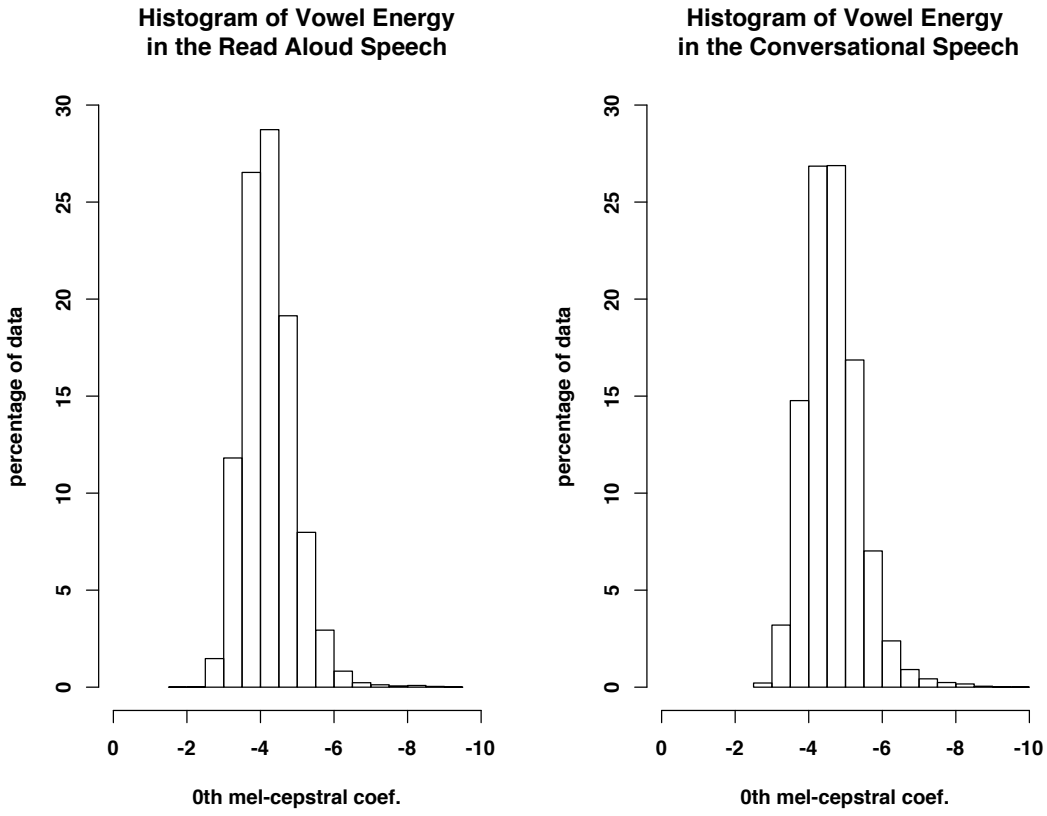

Figure 3.6: Histogram of energy, measured as the 0th mel-cepstrum coefficient, in the centre of the vowels in the read aloud and conversational speech.

Heather	Roger
sil_um_sp	sil_um_sp
sp_I_think	sp_um_sil
sp_um_sp	sp_um_sp
sp_like_sp	sil_but_sp
sil_yeah_sp	I_think_sp
kind_of_sp	sp_you_know
sp_but_sp	sil_I_mean
sil_I_think	sp_I_mean
um_sp_I	um_sp_I
sp_and_sp	but_sp_um
sp_uh_sp	I_mean_I
sp_you_know	sil_and_I
sp_kind_of	sil_yes_I
I_think_sp	sp_uh_sil
but_sp_I	and_I_think

Table 3.5: Examples of trigrams occurring five times or more in Heather's and Roger's conversational speech. Including utterance beginning/end as "sil", and utterance internal pauses as "sp".

Although the read aloud and conversational speech were recorded in the same studio using the same microphone, they were recorded at different times by two different engineers. The similar distributions of vowel energy in figure 3.6 show that we managed to keep recording levels fairly consistent and that there is no substantial difference between the conversational and read aloud data.

3.5.2.2 Fundamental Frequency

The fundamental frequency ($F0$) of the read aloud and conversational speech data reported in this section was extracted for all the speech data when building the HMM-based voices described in section 4.2.

Figure 3.7 shows that Johnny had approximately the same pitch range and $F0$ distribution when reading aloud isolated sentences and when speaking in a conversation. However, Figure 3.7 also shows that the conversational speech had more variation in utterance final $F0$, probably because of more variation in dialogue acts (questions, con-

vowel	i	ɪ	ɛ	æ	ɑ	
genre	read spon	read spon	read spon	read spon	read spon	
mean	18.7 22.6	15.1 17.1	16.9 18.2	17.0 21.1	17.4 16.3	
sd	8.4 6.6	5.7 5.3	4.8 5.4	3.5 6.9	9.1 6.3	
vowel	ʌ	ɔ	o	ʊ	u	filled
genre	read spon	read spon	read spon	read spon	read spon	pauses
mean	18.0 18.2	15.8 14.1	13.6 17.2	13.7 15.9	14.9 14.7	24.4
sd	8.1 6.3	10.0 8.7	7.0 7.7	5.6 8.5	6.9 6.4	7.8

Table 3.6: Spectral tilt: *H1*-A3** measured in decibel (dB).

firmations, etc.) and speaker state (enthusiastic, doubtful, polite, etc.). But mainly, the lack of utterance final $F0$ variation in the read aloud data, like the lack of variation in speaking rate in Figure 3.8, points out the consistency of the task of carefully reading aloud isolated sentences compared to speaking spontaneously in a conversation.

3.5.2.3 Duration and Speaking Rate

The speaking rate, shown in Figure 3.8, of the conversational and read aloud data was measured for speech sequences delimited with silent pauses, as syllables per second. The variation in length of utterances was larger in the conversational data. In order to remove effects of very short and very long utterances, the speaking rate was only measured for utterances that were five to ten words long.

Figure 3.9 shows the duration of the monophthong vowels in the read aloud and the conversational speech. In general the median duration of the read aloud vowels was higher than in the conversational speech, except in the /ʌ/ vowel, because it contained the filled pauses in the conversational speech.

The conclusion drawn was that reading prompts presented in isolation gives a very consistent speaking rate compared to speaking in a conversation.

3.5.2.4 Spectral Tilt

The spectral tilt measure that was used, *H1*-A3**, was described in Hanson (1997); Hanson and Chuang (1999). The measure shows the difference in amplitude between the first harmonic (H1) and strongest harmonic in the third formant (A3), measured in decibel (dB). The measure was chosen because it shows properties of the vocal fold vibration that are related to voice quality (Hanson, 1997; Hanson and Chuang, 1999).

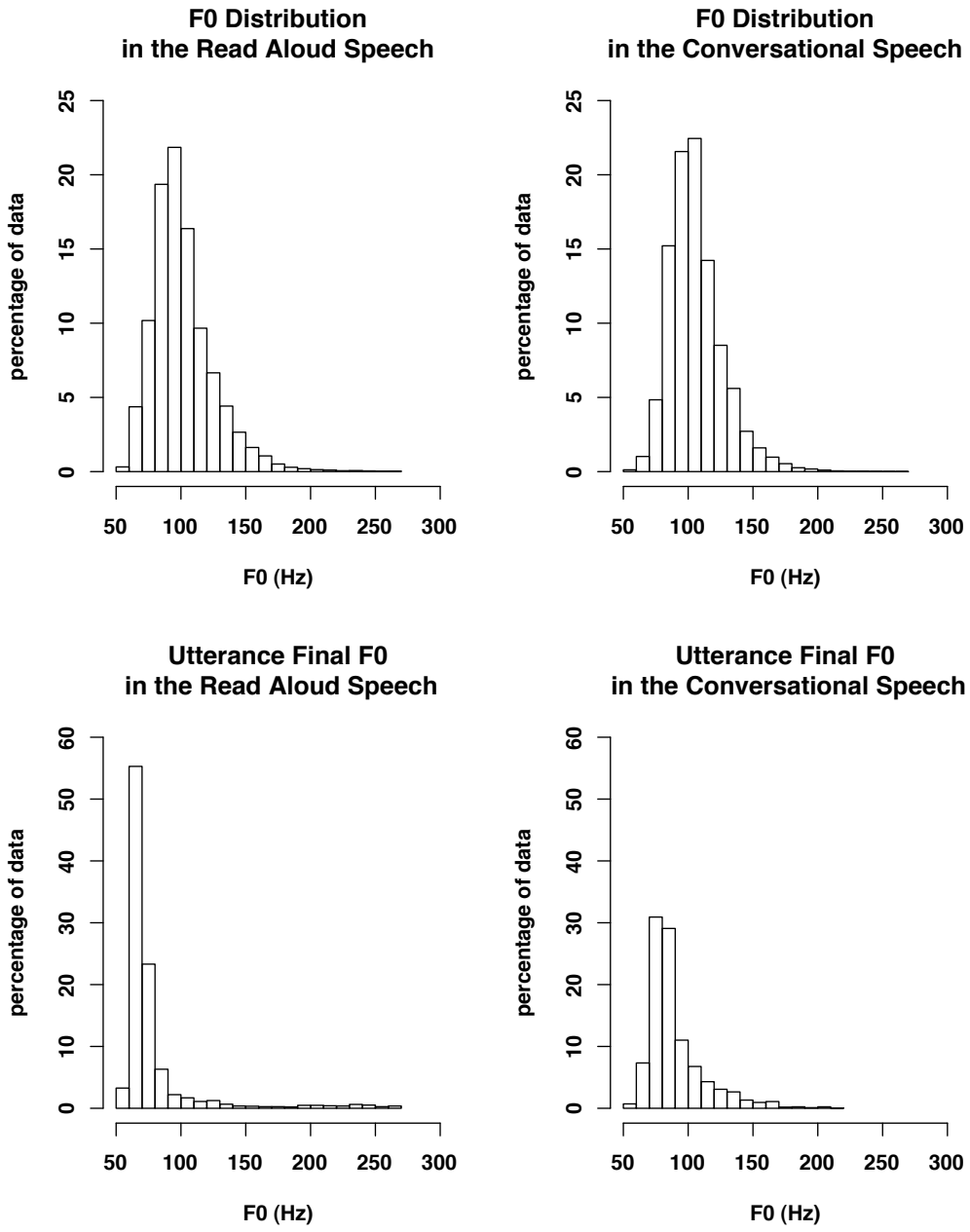

Figure 3.7: $F0$ distribution in read aloud and conversational speech. The top panes show histograms of $F0$ distribution of all voiced frames in the speech data. The bottom panes show histograms of utterance final $F0$ distribution. Due to uncertainties of $F0$ at the end of utterances, the utterance final $F0$ was measured at the tenth last voiced frame, frame length was 5ms.

The spectral tilt was estimated using scripts written by Timothy Mills[3]. The result is presented in Table 3.6. The *H1*-A3** measure relies on correctly estimated $F0$ and

[3]http://nuweb.neu.edu/tmills/scripts.html

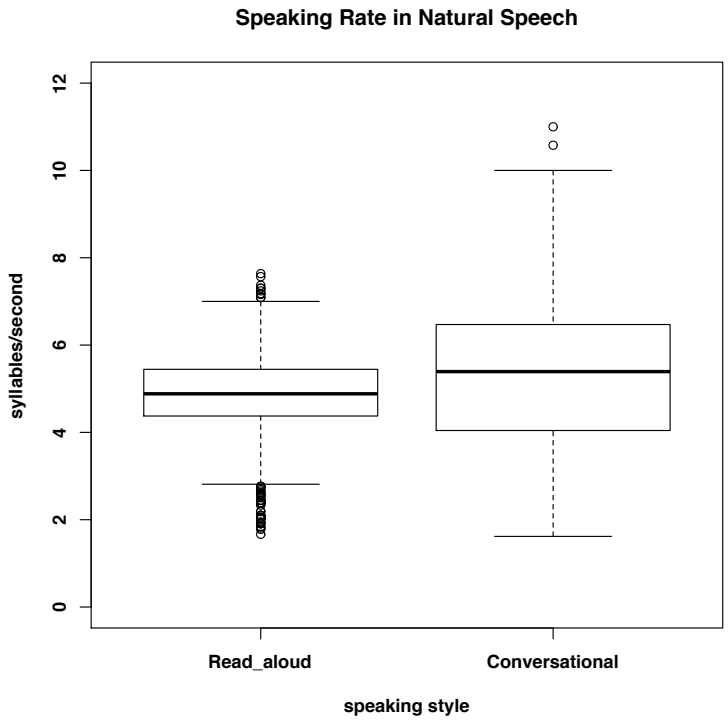

Figure 3.8: Speaking rate for utterances with 5-10 words in the conversational and read aloud data. The solid line is the median, box borders show the upper and lower quartiles, and the whiskers are drawn to 1.5 times the inter-quartile range (IQR).

the first three formants and required manual supervision. Therefore, the spectral tilt was only measured for vowels that fulfilled certain criteria: lexically stressed vowel, with at least a median duration (but not longer than $1.5 * IQR$, see figure 3.9), from a content word. Only one vowel of each type was extracted from an utterance, e.g. not two /ʌ/ from the same utterance, to increase the spread of sampled utterances across the data. The fifty first samples from each monophthong, except the two schwas (/ə/ and /ɚ/), that fulfilled these criteria were selected from the speech data. This selection gave a total of five hundred vowels from the read aloud speech and five hundred and fifty (including filled pauses as a separate vowel type) vowels from the conversational speech. The selection criteria ensured that the extracted vowels spanned across a minimum of a few hundred utterances for the majority of the vowels up to about a thousand utterances for the vowel /ʊ/ in both the read aloud and conversational speech.

The script extracted $F0$ and formants from the centre of the forced aligned vowels. Manual adjustments of window position for formant extraction and manual mark up of missing or erroneous pitch periods, allowed reliable estimates for almost all the

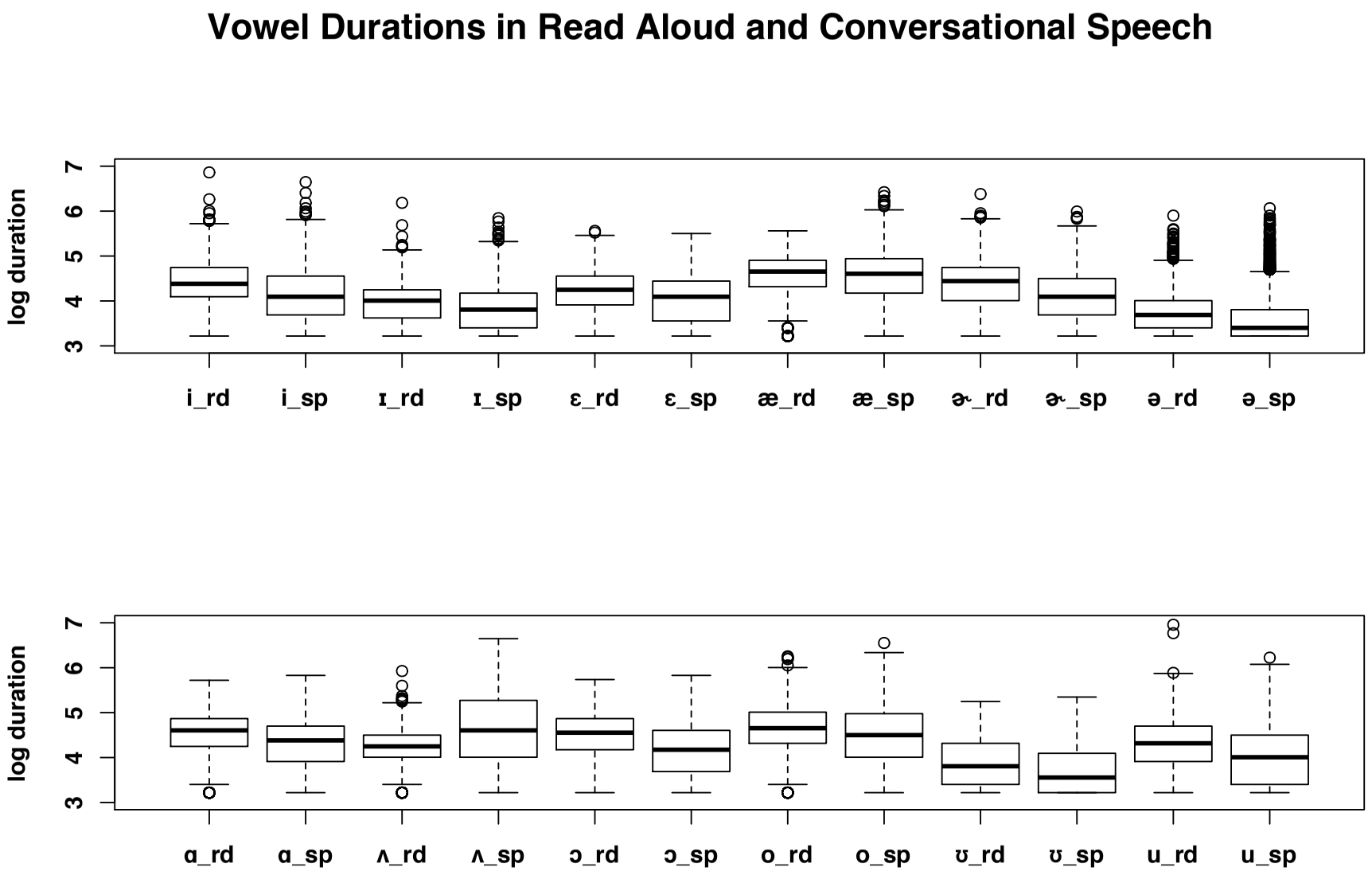

Figure 3.9: Vowel duration, measured in log milliseconds, in the read aloud (*_rd) and conversational (*_sp) speech data. The solid line is the median, box borders show the upper and lower quartiles, and the whiskers are drawn to 1.5 times the inter-quartile range (IQR).

vowels. A few samples were however discarded: In the /ɔ/ vowel the first two formants lie very close and could in five instances each in the read aloud and conversational speech not be reliably estimated and were discarded, which for this vowel left forty-five instead of fifty tokens for spectral tilt analysis. One instance of the vowel /ʊ/ in the conversational speech was discarded due to that no pitch period could be reliably identified.

The speech data in Hanson (1997); Hanson and Chuang (1999) came from males and females reading aloud carrier phrases and the average values of *H1*-H3** was 13.8 dB for male speakers and 23.4 dB for female speakers, and standard deviation was 4.8 dB and 6.6 dB, respectively (Hanson and Chuang, 1999). Johnny's averages for each vowel in table 3.6 are inbetween those values and show that the majority of the read aloud and conversational speech were spoken with a modal voice quality.

3.5.2.5 Vowel Quality

Figure 3.10 shows the average frequencies of automatically extracted first and second formants from the centre of the vowels in the forced aligned speech data. The extracted values were verified manually by spot checking data from some of the vowel types where the average formant values deviated markedly from more prototypical formant values. The prototypical formant values were taken from Ladefoged (2006) who gives: "[the] average of a number of authorities' values of the frequencies of the first three formants in eight American English vowels."(Ladefoged, 2006, p.184). The manual check confirmed the average values for the front vowels, for example /i/ and /æ/, but revealed problems with some of the back vowels. The small distance between F1 and F2 in /ɔ/ made the extracted values unreliable. In the /u/ vowel the formants were not well estimated for the word *you* because, due to coarticulation, F2 starts off high. The vowel in the filled pauses were stipulated in the lexicon as an /ʌ/, but as shown in figure 3.10, this was not entirely correct and was the main reason for the difference between the averages for the /ʌ/ vowel in the read aloud and conversational speech.

The read aloud and conversational speech both contained a large proportion of unstressed and unaccented syllables and a generally reduced vowel space is to be expected. The spectral tilt in Section 3.5.2.4 relied on estimated formant values that were measured for lexically stressed vowels in content words with at least a median duration. These formant values, shown in figure 3.11, were expected to be less centralised than in figure 3.10 and because each vowel was manually verified and taken from a more restricted phonetic context they should also be able to better capture differences

in vowel quality between the read aloud and the conversational speech.

Compared to formant frequencies taken from Ladefoged (2006), which show idealised or prototypical formant frequencies, figure 3.10 shows a more reduced vowel space than figure 3.11 for both the read aloud and conversational speech. Contrary to Nakamura et al. (2008), neither figure 3.10 nor figure 3.11 show an obvious tendency to a reduced vowel space for the conversational speech compared to the read aloud speech. Part of the explanation for this was the careful selection of the conversational speech and the automatic assignment of schwa in reduced pronunciations, but the result also confirms our intuition that our speaker Johnny did not have a particularly enunciated reading style. However, there were some observable reduction tendencies, and an example of differences in vowel formant values in fully pronounced and reduced *it* is shown in figure 3.12. Bell et al. (2003) showed that utterance initial vowels were more likely to be fully pronounced, and the difference between *it* in the read aloud and conversational data is likely to be due to the distributional differences in utterance position. In the read aloud data 77 out of the 193 (40%) *it* occur in utterance initial position, whereas only 2 out of the 291 *it* in the conversational data occur in utterance initial position. It is possible that we could find other words like *it* where the vowel quality differed between the conversational and read aloud data due to differing phonetic context, but the analysis in figure 3.10 and figure 3.11 does not support a general difference in vowel quality between read aloud and conversational speech that would prevent blending them in speech synthesis.

3.6 Conclusion

In this chapter we showed how a spontaneous conversation was recorded, transcribed and analysed. The purpose was to obtain conversational speech suitable for building unit selection and HMM-based synthetic voices.

It may be possible to make synthetic speech exhibit phonetic properties similar to conversational speech without the use of actual conversational speech data or with other speech synthesis methods than unit selection or HMM-based speech synthesis, but as we stated in section 1.1:

- Unit selection and HMM-based speech synthesis are currently the two dominating frameworks due to their ability to build high quality synthetic voices by mimicing the speech properties from recordings of natural speech (see e.g. King

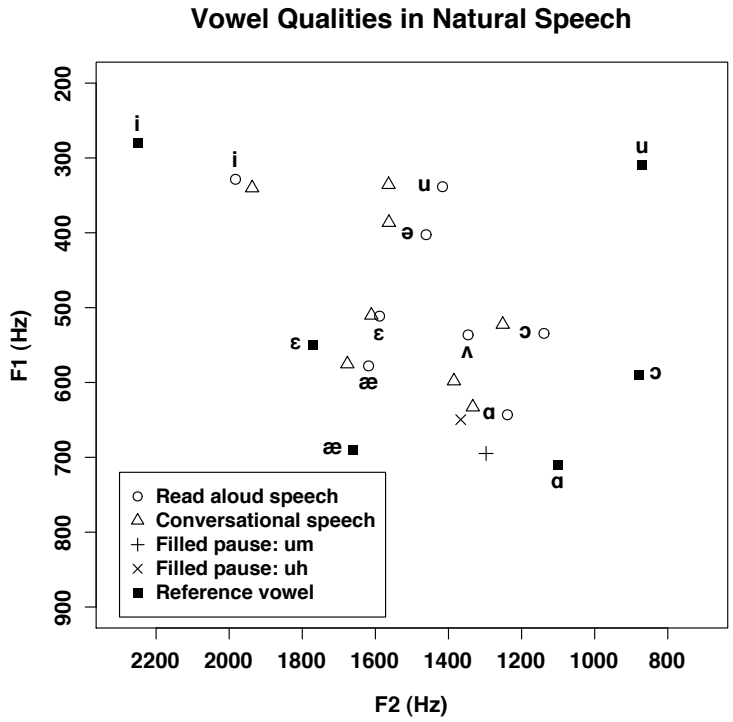

Figure 3.10: Mean formant values (F1 and F2) for American English monophthongs in the read aloud and the conversational speech. The reference formant values are taken from Ladefoged (2006). The mean formant values for the two filled pause types (*um* and *uh*) in the conversational speech are also plotted.

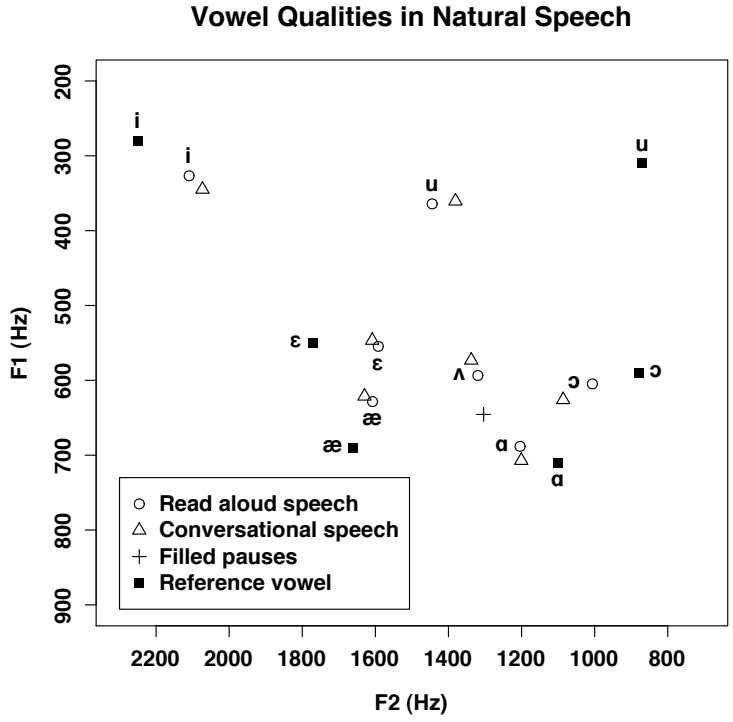

Figure 3.11: Mean formant values (F1 and F2) for manually checked American English monophthongs from lexically stressed content words in the read aloud and the conversational speech. The reference formant values are taken from Ladefoged (2006). The mean formant value for the filled pauses (*um* and *uh*) in the conversational speech is also plotted.

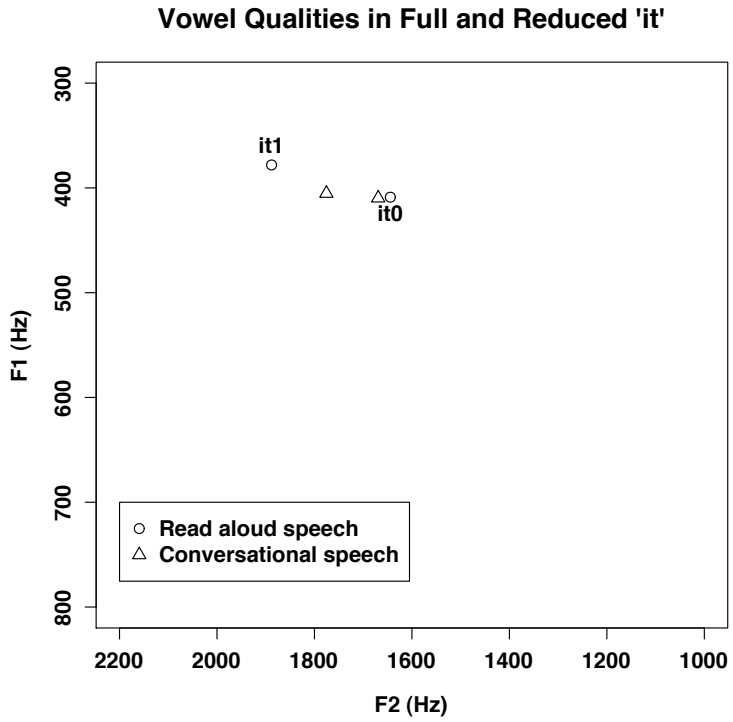

Figure 3.12: Mean formant values (F1 and F2) for fully pronounced and reduced vowel in the word *it*. Represented as it1 (full) and it0 (reduced) in the figure.

and Karaiskos, 2010). Building synthetic voices from conversational speech data should therefore result in synthetic voices with conversational characteristics.

- Any perceived spontaneity from speech that is not spontaneous speech will be determined by the quality of the actor (see section 1.1). Using speech from a spontaneous conversation therefore allows the focus of our work to be put on whether the synthesis and evaluation methods are appropriate for developing conversational speech synthesis, rather than if the actor is good enough.

The conversational speech phenomena described in section 2.2 and the description of a recorded conversation in this chapter, suggested that using conversational speech data in unit selection and HMM-based speech synthesis systems currently represents the most feasible method for adding conversational characteristics to synthetic voices. This is the approach tested in this thesis.

Eliciting conversation in a recording studio proved to be a straightforward method for obtaining speech that contained a rich variety of spontaneous conversational speech phenomena. Section 3.5 showed that a large proportion of the recorded conversation consisted of the discourse markers, filled pauses, and backchannels that were described

in section 2.2.

The recorded conversation was transcribed manually (see section 3.4), but all subsequent processing of the speech was performed automatically:

- Section 3.4.5 showed that the conversational speech could be segmented by adjusting the forced alignment method in our speech synthesis system.
- Propositional content is generally represented in speech synthesis through linguistic features such as neighbouring phonemes, and position of syllable in word and utterance. In section 2.2 we argued that these features would suffice to preserve also the phonetic properties and pragmatic functions of conversational speech phenomena, such as discourse markers and filled pauses. The frequency of the discourse markers and filled pauses together with their local phonetic context shown in table 3.4 should allow them to be selected from an appropriate context in unit selection, as well as allow capturing their phonetic properties in the training of HMM-based synthetic voices.

Hence, synthetic voices that sound like a person participating in a spontaneous conversation can be built from conversational data with conventional unit selection and HMM-based speech synthesis systems. However, figure 3.4 shows that the lack of control over the phonetic material in conversational speech makes it problematic to achieve phonetic coverage. This lack of coverage and the formulations of the unit selection and HMM-based speech synthesis frameworks makes it challenging to synthesise consistently high quality utterances that are not pre-recorded.

The phonetic analysis of the read aloud and conversational speech data in section 3.5.2 showed that the only general differences between the conversational and read aloud speech were the speaking rate and vowel duration. Other differences found were related to the local context, such as the phonetic properties of the filled pauses. This suggested that the important differences for conversational speech synthesis are in the local phonetic properties of specific speech phenomena, in particular the differences related to the language composition in section 3.5.1.

The approach taken in this thesis was therefore to alleviate the lack of phonetic coverage in the conversation by blending it with a conventional speech resource of pre-selected and read aloud sentences. The blending required taking into account the differences in language composition and phonetic properties of the conversational and read aloud data. In chapter 4 we will describe our developed blending techniques for unit selection and HMM-based speech synthesis, and in chapter 5 we will describe the

perceptual evaluations of our blended conversational synthetic voices. We will demonstrate that conversational speech data and blending can be successfully used to build synthetic voices with richer behaviour than conventional voices. The conversational data allowed us to synthesise natural-sounding conversational characteristics, in particular discourse markers and filled pauses. The added read aloud data allowed us to fill in the gaps in phonetic coverage and synthesise also high quality propositional content. This allowed our synthetic voice to express certainty and uncertainty about a topic in a manner similar to how humans express it in spontaneous conversation.

Chapter 4

Synthetic Voices

This chapter will describe how the HMM-based and unit selection voices were built. The chapter includes descriptions of initial attempts of utilising spontaneous conversational speech for unit selection and HMM-based speech synthesis (see sections 4.2.3 and 4.3.2). These voices were built with small amounts (approximately 20min) of conversational speech from Heather and Roger, because at the time we did not have the larger amount of speech from Johnny.

Both the positive and negative results from those pilots were the motivating factor behind the recordings of a larger amount of spontaneous conversation from Johnny, described in chapter 3, and the final unit selection and HMM-based synthetic voices described in this chapter.

The names for the synthetic voices we describe in this chapter, e.g. joh.16k.hts.read, follow the following naming convention: the first part is a three letter abbreviation of the speaker (e.g. joh stands for Johnny), the second part is the sampling rate of the speech data (e.g. 16k for 16kHz), the third part stands for the type of system used (hts for HTS and unit for unit selection), and the last part contains additional info on the type of speech data or synthesis technique used in the voice (e.g. read when the voice contains only read aloud data, or blend when we use our blending techniques in voices that contain both read aloud and conversational data).

4.1 Independent Contribution by the Author: Design and Analysis of the Synthetic Voices

The synthetic voices described in this chapter were part of joint work in Andersson et al. (2010a), Andersson et al. (2010b) and Andersson et al. (2012). This section outlines the author's independent contribution to the design, building and analysis of the synthetic voices in this chapter:

- All preparation of speech data and building of the blended unit selection and HMM-based voices were carried out by the author of this thesis.
- The design and implementation of the unit selection and HMM-based blending methods in sections 4.2.5, 4.3.3, and 4.3.4 were all made by the author of this thesis.
- All the analysis of the synthetic speech in sections 4.2.6 and 4.3.6 was made by the author of this thesis. The majority of these analyses were not part of the joint publications.

In general, all reported work was made by the author of this thesis, unless explicitly stated otherwise. For example: two of the reference voices used in the pilots in sections 4.2.3 and 4.3.2 are credited to other people.

4.2 HMM-based Voices

The HMM-based voices described in sections 4.2.4 and 4.2.5 were used in Andersson et al. (2010b) and Andersson et al. (2012). These voices were built by the author using scripts provided by Junichi Yamagishi. The scripts were modified by the author for the blended voices described in Section 4.2.5.

4.2.1 The Context-dependent Phonemes

The HTS system does not include text analysis and the generation of the context-dependent phonemes. The context-dependent phonemes were therefore generated with the CereVoice system from the text and speech analysis used for the unit selection voices (see section 4.3). CereVoice's contexts were based on the contexts in Tokuda et al. (2002); Zen et al. (2007) and its more recent variant in Zen et al. (2009), and took into account:

- quinphone (i.e. current phoneme with the two preceding and succeeding phonemes as context, example: s-p-ɔ-r-t)
- preceding, current, and succeeding phoneme types (vowel, plosive, etc.)
- nucleus of current syllable (e.g. æ, ɔ or ʌ)
- position of phoneme in syllable, word and phrase
- position of syllable in word and phrase
- number of phonemes in syllable, word and phrase
- number of syllables in word and phrase
- part-of-speech (content or function word)
- preceding, current, and succeeding syllable stress and accent
- boundary tone of phrase (utterance final or -medial)

The contexts did not include explicit representations of the discourse markers or filled pauses (*um* or *uh*), but the context specifications implicitly identified many important characteristics. The quinphone context was large enough to encapsulate many of the discourse markers and filled pauses, e.g. *yeah*, *you know* or *oh yeah*, together with their, often initial or final, utterance positions (see table 3.4). The quinphone context was also large enough to include the filled pauses together with a preceding short function word, such as *and* or *but*, or a common word ending, such as *-ing*, and thereby potentially preserving any associated preceding hesitation. The contexts with counts and phrase positions should also be able to capture segmental and prosodic differences between the same word token in different utterance contexts, as in the previously mentioned example in section 3.4 of *yeah* as a stand alone backchannel, in the confirmation *yeah yeah yeah*, or in the longer utterance *yeah I feel kind of dirty afterwards*.

Our hypothesis was that the current context representations would be sufficient to build HTS voices where the contrast between different data sources; conversational or read aloud speech data, could be preserved. The result would be that voices including conversational speech would generate more natural-sounding conversational speech phenomena, such as discourse markers and filled pauses.

4.2.2 Building HTS Voices

The HTS toolkit[1] with which the voices were built is a patch to the HTK speech recognition toolkit (Young et al., 2006). The method and training scripts used to build the

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

HMM-based voices were developed by Junichi Yamagishi. The scripts follow the general methodology of the HTS system (Zen et al., 2007) that was described in section 2.3.3. The training procedure was the same for all the HMM-based voices in this thesis.

The speech samples were downsampled from 48kHz to 16kHz. Spectral and excitation parameters were extracted from the speech samples with 25ms window and 5ms frameshift as 39th order STRAIGHT (Kawahara et al., 1999) mel-cepstrals, five frequency band averaged aperiodicity (Kawahara et al., 2001), $logF0$, together with their delta, and delta-delta values.

Gaussian distributions of the acoustic parameters, and duration, were then trained for the context-dependent phonemes described in section 4.2.1. The context-dependent phonemes were represented as 5-state left-to-right Hidden Markov models (HMMs), where the acoustic parameters were trained as five independent streams (one stream each for mel-cepstral, aperiodicity together with their delta and delta-delta values, and three separate streams for $logF0$, delta $logF0$ and delta-delta $logF0$).

The training of the context-dependent models follows largely the training of HMM models for speech recognition as outlined in Young et al. (2006), but with extensions to allow for modelling voiced and unvoiced sequences of speech (Tokuda et al., 1999) and a better representation of duration for speech generation (Zen et al., 2004b), resulting in the HTS specific MSD-HSMM modelling. Firstly, the context is stripped from the context-dependent phonemes and monophone HMM models, one for each phoneme, are trained to obtain robust initialisation values for the context-dependent models. Secondly, the monophone models are converted back into full context models and trained with embedded training with maximum likelihood criterion. Thirdly, the large context gives few instances of each context-dependent phoneme type, and when synthesising speech, models which are not in the training data need to be dealt with. Therefore the parameters are shared ("tied") between the states of the different context-dependent models. The method, decision tree-based context clustering, used to share the model parameters and deal with unseen models when synthesising speech was developed by Odell (1995). The decision tree splits the data into a binary tree, based on the individual contexts in the full context models. The leaf-nodes in the tree contain the trained Gaussian distributions. The decision to stop splitting into more leaf-nodes is determined by the minimum description length (MDL) criterion. A decision tree is created for each of the mel-cepstral, aperiodicity, $logF0$ and duration parameters. To further improve the estimation of the parameters, the process is repeated: the clustered

parameters are "untied" and the full context models are again trained with embedded training, and again clustered into decision trees. The resulting trained models can then be used to generate high quality synthetic speech.

4.2.2.1 Speech Generation

The script for speech generation was developed by Junichi Yamagishi, and was used unmodified for all HMM-based voices in this thesis. Just as for the training, the context-dependent phoneme descriptions for speech generation were generated with the CereVoice system. Speech parameters are then generated from the corresponding trained models in the clustered mel-cepstral, aperiodicity, $logF0$ and duration trees as described in section 2.3.3.4. Firstly, the state model sequence is determined by maximum likelihood generation, giving the mean duration of each model. Then, the spectral and excitation parameters are generated with the speech parameter generation technique that considers the global variance (Toda and Tokuda, 2007), to ensure that the generated utterance has a smooth trajectory with natural variation.

4.2.3 Pilot: Read Aloud to Spontaneous Adaptation

Initial experiments with utilising spontaneous speech for HTS revealed that adapting read aloud voices with spontaneous speech did not result in perceptually favourable distinctions of speaking styles.

The data described in section 3.2.1 had only about 300 utterances with 22min of phonetic material for the male speaker, Roger, which was not enough to build a good quality speaker-dependent voice from. Instead we utilised the adaptation technique described in Yamagishi et al. (2007). The aim was to adapt an existing read aloud voice into a voice with a more spontaneous speaking style, by using a small amount of spontaneous speech as adaptation data. The read aloud source voice, henceforth rog.16k.hts.read, was built from several hours of neutrally read aloud sentences from Roger. This voice was built by J. Yamagishi with the HTS system configurations described in Zen et al. (2007). The rog.16k.hts.read voice was then adapted with the 22min of spontaneous speech from Roger. This adapted "spontaneous" voice is henceforth referred to as rog.16k.hts.adapt.

The adaptation data was automatically forced aligned and converted into Festival's utterance structure using the tools in Clark et al. (2007). The conversion into context-dependent phonemes, as well as the adaptation itself, was made using scripts from

Yamagishi et al. (2007).

4.2.3.1 Perceptual Evaluation

The test sentences were taken from held-out conversational material from Roger. These nineteen held-out utterances, shown in table A.1 in appendix A, were synthesised with the spontaneous (rog.16k.hts.adapt) and read aloud (rog.16k.hts.read) voices. The utterances were presented in pairs to volunteering listeners. The listeners were asked to judge which utterance in the pair had *the most spontaneous speech quality* and which had *the best general speech quality*[2]*, regardless if it sounds spontaneous or not* or if they were equal in any of these aspects. The order of the speech between and within pairs was randomised between listener. Twenty-two listeners, both native and non-native English speakers, took part in the evaluation.

4.2.3.2 Results

The perceptual judgements have been collapsed over all utterances and are shown in figure 4.1. The significance of the result was tested with the binomial test. The number of times participants judged the quality as equal ("No preference") was removed before testing the significance.

The spontaneous quality was not generally perceived in the adapted voice ($p = 0.74$) and the original read aloud voice had a significantly higher ($p < 0.001$) general speech quality.

4.2.3.3 Discussion and Conclusions

When adapting a read aloud speaking style to a spontaneous speaking style the prosodic categories and dependencies are determined by the read aloud voice. The topology of the decision trees of the read aloud voice are fixed and only the parameters of the Gaussian distributions at nodes in the trees are adapted. This means that segmental and prosodic categories and dependencies that are not represented in the original voice will not be present in the adapted voice. This resulted in inappropriate realisations of spontaneous speech phenomena, like the filled pauses, since they do not exist in the read aloud voice. There are likely to also be other problems of style adaptation that

[2]In the pilot experiments in sections 4.2.3 and 4.3.2 "general speech quality" was used instead of the more conventional formulation of how "natural" an utterance sounds. We assumed that naturalness and general quality measured very similar aspects, but we have not shown that, and therefore we used the conventional formulation with naturalness in the experiments in chapter 5.

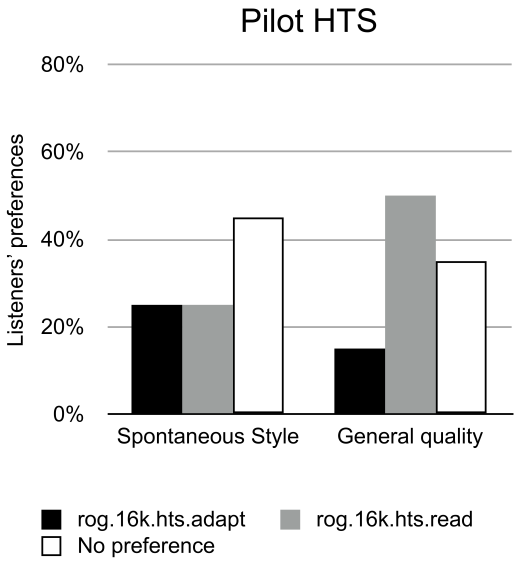

Figure 4.1: Raw data table. Perceptual judgements of spontaneous and general speech quality in HMM voices adapted with spontaneous speech (rog.16k.hts.adapt) or trained from read aloud speech (rog.16k.hts.read). The “No preference” shows the proportion of listeners who expressed no preference for any of the voices.

contributed to the generally lower quality of the adapted voice compared to the original voice. Based on the lack of perceived spontaneity and the lower quality of the adapted voice, we will focus on the speaker dependent HMM-based speech synthesis as it is described in Zen et al. (2007) and section 4.2.2. In the work with speaker dependent voices we will utilise the larger Johnny data described in chapter 3 and summarised in table 3.2.

4.2.4 Conversational and Read Aloud HTS Voices

The context-dependent phonemes in section 4.2.1 for the read aloud and conversational speech were used to build one “spontaneous” and one “read aloud” synthetic voice with the method described in section 4.2.2. The data used was described in chapter 3 and a summary is provided in table 3.2. These voices are henceforth referred to as joh.16k.hts.spon and joh.16k.hts.read, respectively.

The sizes of the clustered decision trees reflect the amount and complexity of the speech data. Table 4.1 shows that the mel-cepstral, aperiodicity and $\log F0$ trees were smaller in the joh.16k.hts.spon voice than in the the joh.16k.hts.read voice, due to less data and less phonetic coverage. The duration trees were however almost equally large due to more variation in duration in the conversational data.

Table 4.1: Number of leaf nodes in the clustered duration, $\log F0$, mel-cepstral and aperiodicity trees, for the joh.16k.hts.spon (SP) and joh.16k.hts.read (RD) voices. The ratio(SP/RD) shows the relative tree sizes.

	SP	RD	Ratio (SP/RD)
duration	1699	1602	1.06
$\log F0$	4618	5248	0.88
mel-cepstral	837	1405	0.60
aperiodicity	994	1543	0.64

4.2.5 Blending Read Aloud and Conversational Speech

Our first impression of the quality of the joh.16k.hts.spon voice was that whereas the discourse markers and filled pauses could be synthesised with quite high quality, the quality of the propositional content was often less good. To increase the phonetic coverage, and thereby improve general segmental and prosodic quality, while still preserving important conversational characteristics, all the conversational and read aloud data in table 3.2 were pooled in the training and clustering of HMM-based models. An additional context: speaking style (spontaneous or read), was added to the context-dependent phoneme descriptions in section 4.2.1. The method to represent the different data sets with a style context has previously been successfully applied to blend and preserve different "emotional" speaking styles (Yamagishi et al., 2005).

When training the context-dependent HMM-based models, the speaking style context was then available as a question in the decision tree based clustering. The speaking style context was automatically selected as an important feature throughout the clustering process. For example, in the duration clustering, the speaking style context was selected almost immediately to split the data based on the difference in duration of the syllable nucleus between the conversational and read aloud speech. For the excitation and spectral part, the sharing or splitting based on the speaking style context was more complex.

During synthesis with this voice one of the speaking styles was selected for an utterance by setting the speaking style context to either spontaneous or read aloud for all context-dependent phonemes, and then speech parameters were generated. Henceforth, utterances generated in this way are referred to as from the joh.16k.hts.blendspon voice and joh.16k.hts.blendread voice, respectively.

4.2.6 Phonetic Properties of the Synthetic Speech

A test set of synthetic speech was generated from each of the synthetic voices: the joh.16k.hts.spon, the joh.16k.hts.read, the joh.16k.hts.blendspon and the joh.16k.hts-.blendread voice. The context-dependent phonemes for the synthetic speech in the test set were obtained from unused transcripts of Johnny's speech. The benefit of using this material as test sentences was that it was from the same speaker as in the training data, hence representing his way of expressing himself. The material was rich in conversational speech phenomena with nearly one hundred filled pauses, eighty-one *yeah* and at least a few instances each of e.g. *okay*, *right* and *oh*.

This gave us a set of 169 utterances for each synthetic voice that was rich in conversational phenomena, and had identical phonemic sequences and linguistic analysis, thus allowing a linguistically balanced acoustic comparison.

In section 3.5.2 we showed comparisons of segmental and prosodic properties in the read aloud and conversational data. In this section we will show segmental and prosodic properties of the synthetic voices built with either conversational or read aloud speech, and the blended voice built with both.

Figure 4.2 shows a comparison of the first two formants in the synthetic speech. For the joh.16k.hts.spon and the joh.16k.hts.read voices the mean formant values were generally similar to each other, and similar to the natural speech. As in the natural speech, there was no strong tendency towards a reduced vowel space in the joh.16k-.hts.spon compared to the joh.16k.hts.read. The synthetic /u/ vowels were difficult to automatically extract formants from in all the synthetic voices, because, due to coarticulation, F2 starts off high in non-reduced *you*, *to*, *do* and *doing*. Figure 4.2 also shows that the vowel qualities were slightly closer to each other in the blended voice than in the style dependent voices, joh.16k.hts.spon and joh.16k.hts.read, or in the natural read aloud and conversational speech. This pattern was also preserved in the synthetic speaking rate, shown in figure 4.3, where the joh.16k.hts.spon and joh.16k.hts.read preserved the speaking rate differences in the natural speech, but the blending resulted in more similar speaking rates.

On the other hand, both duration and vowel quality of filled pauses in natural conversational speech were to a large extent preserved in the joh.16k.hts.spon, as well as in the joh.16k.hts.blendspon voice, and different from the vowel quality and duration in the joh.16k.hts.read (see figures 4.4 and 4.5). The duration of *um* synthesised with the joh.16k.hts.read voice did not have much similarity to the duration of *um* in the

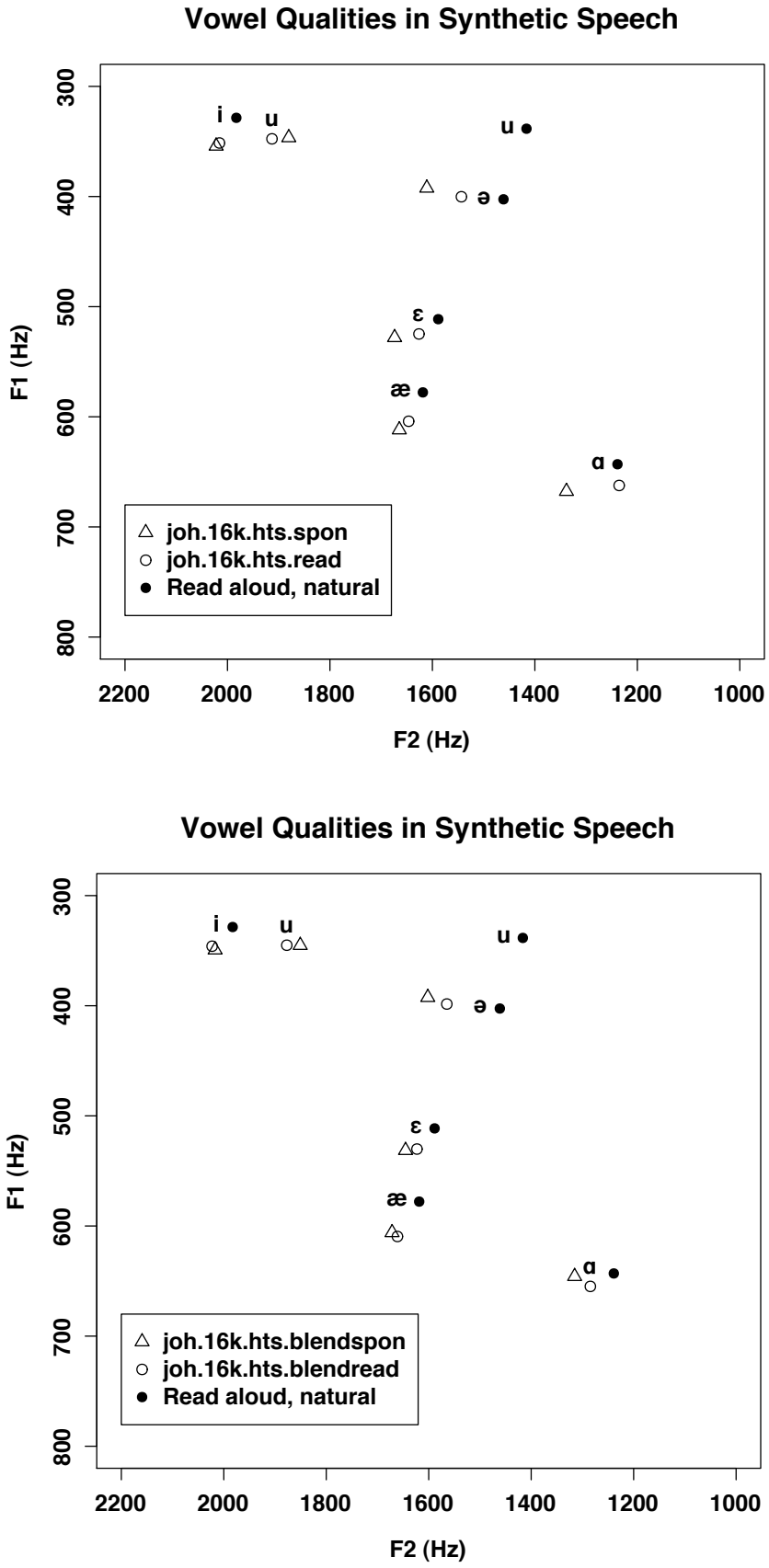

Figure 4.2: Mean formant values (F1 and F2) for American English monophthongs, denoted with IPA symbols, in 169 utterances synthesised with four different synthetic voices. The utterances contained the same phonemic sequences for all four voices. Some of the natural vowels from figure 3.10 are provided as a reference.

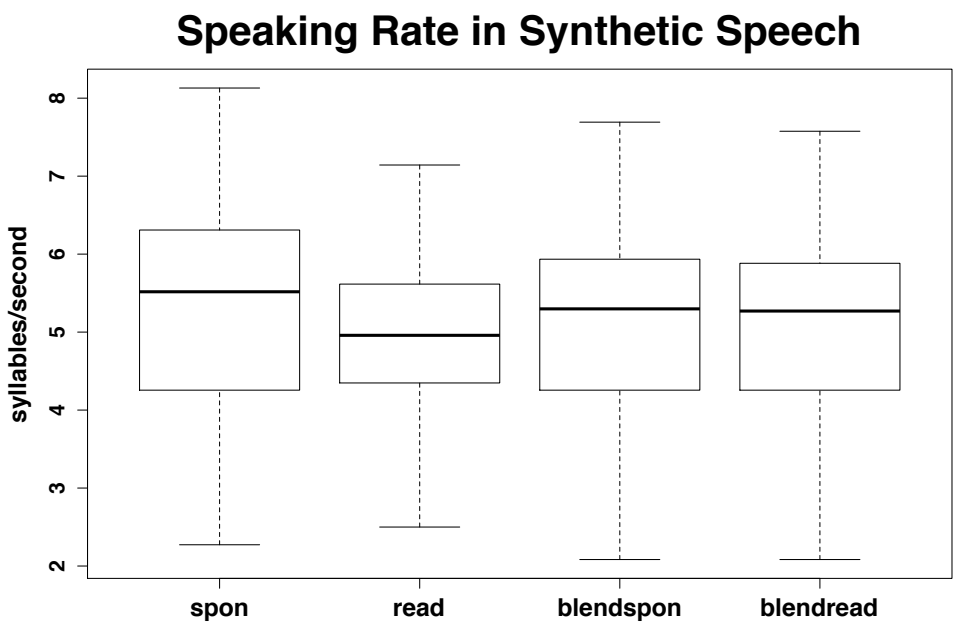

Figure 4.3: Speaking rate for 169 utterances synthesised with four different synthetic voices. The utterances had the same phonemic sequences for all the synthetic voices. The prefix joh.16k.hts is the same in all four voices, and only the suffixes spon, read, blendspon and blendread are written in the figure.

natural speech. The duration of *uh* in the joh.16k.hts.read voice was more similar to the natural filled pauses than the synthetic *um*s from this voice, but there were no filled pauses in the read aloud speech data. The long median duration was due to the long duration of the words *ah* (mean = 260ms) and *oh* (mean = 205ms) in the "conversational style" text in the read aloud coverage material, e.g. in the sentence "Ah well, maybe more next week.". A similar pattern to the filled pauses was also observed for the pitch contour of utterance initial *yeah*, shown in figure 4.6.

In general, there was more variation in the natural speech than in either of the synthetic voices. But, figure 4.8 shows an utterance initial filled pause where the joh.16k-.hts.spon had segmental and prosodic properties similar to a natural reference sample, and hence conveyed a similar degree of hesitation, whereas the segmental and prosodic properties of the *um* from joh.16k.hts.read were different and did not sound much like a filled pause. Similarly, many discourse markers were generally well preserved in both the joh.16k.hts.spon and joh.16k.hts.blendspon. Figure 4.7 shows an utterance initial *yeah*, followed by a short pause, from natural and synthetic speech, where the joh.16k.hts.spon had segmental and prosodic properties similar to the natural reference sample, whereas the *yeah* from the joh.16k.hts.read had different shape of the F0

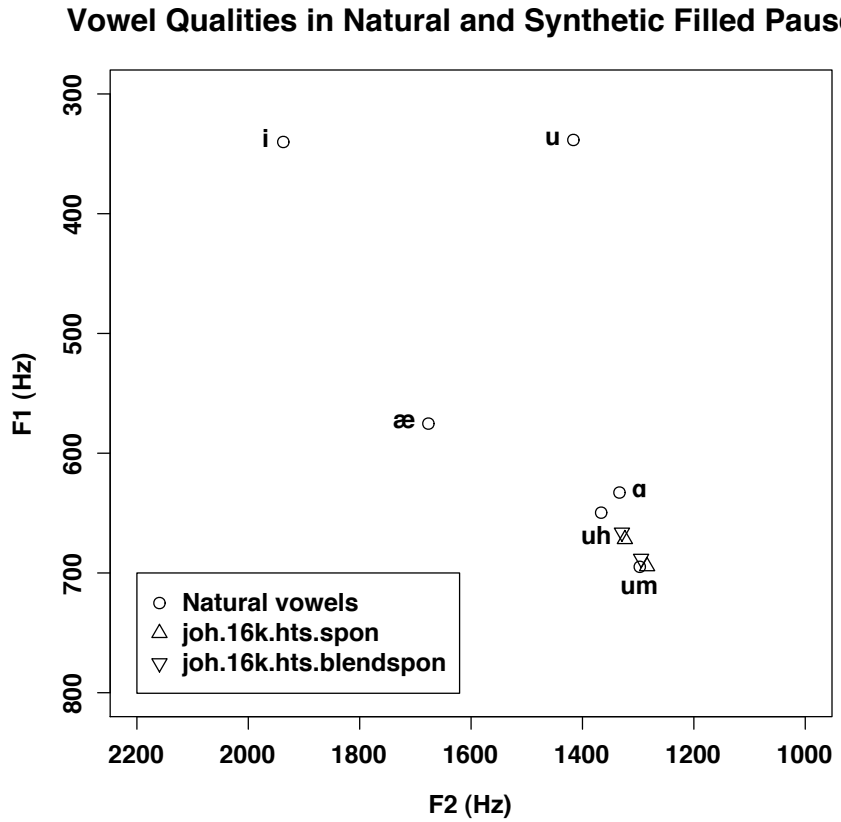

Figure 4.4: Vowel quality of the filled pauses, *um* and *uh*, in conversational speech, joh.16k.hts.spon and joh.16k.hts.blendspon. Natural vowels from the conversational speech are provided to illustrate where in the vowel space the filled pauses lie. (Vowel qualities of the filled pauses in joh.16k.hts.read are not plotted, but they were more different: um F1:661/F2:1342, uh F1:589/F2:1399.)

contour, longer duration of the vowel part of the *yeah*, and despite that the phonemic sequence was intelligible, it came across as almost meaningless.

4.2.7 Alternative Context Representations

In section 4.2.1 we argued that the current shallow context representations of e.g. phoneme sequence and utterance position would be sufficient to generate discourse markers and filled pauses with HMM-based voices. The phonetic analysis in section 4.2.6 supported the use of these shallow representations. However, positive results in Badino et al. (2009) from using alternative context representations to synthesise novel speech phenomena prompted a follow-up investigation.

In Badino et al. (2009) we investigated the inclusion of a novel prosodic category emphasis, in HMM-based synthetic voices. This published work on emphasis mod-

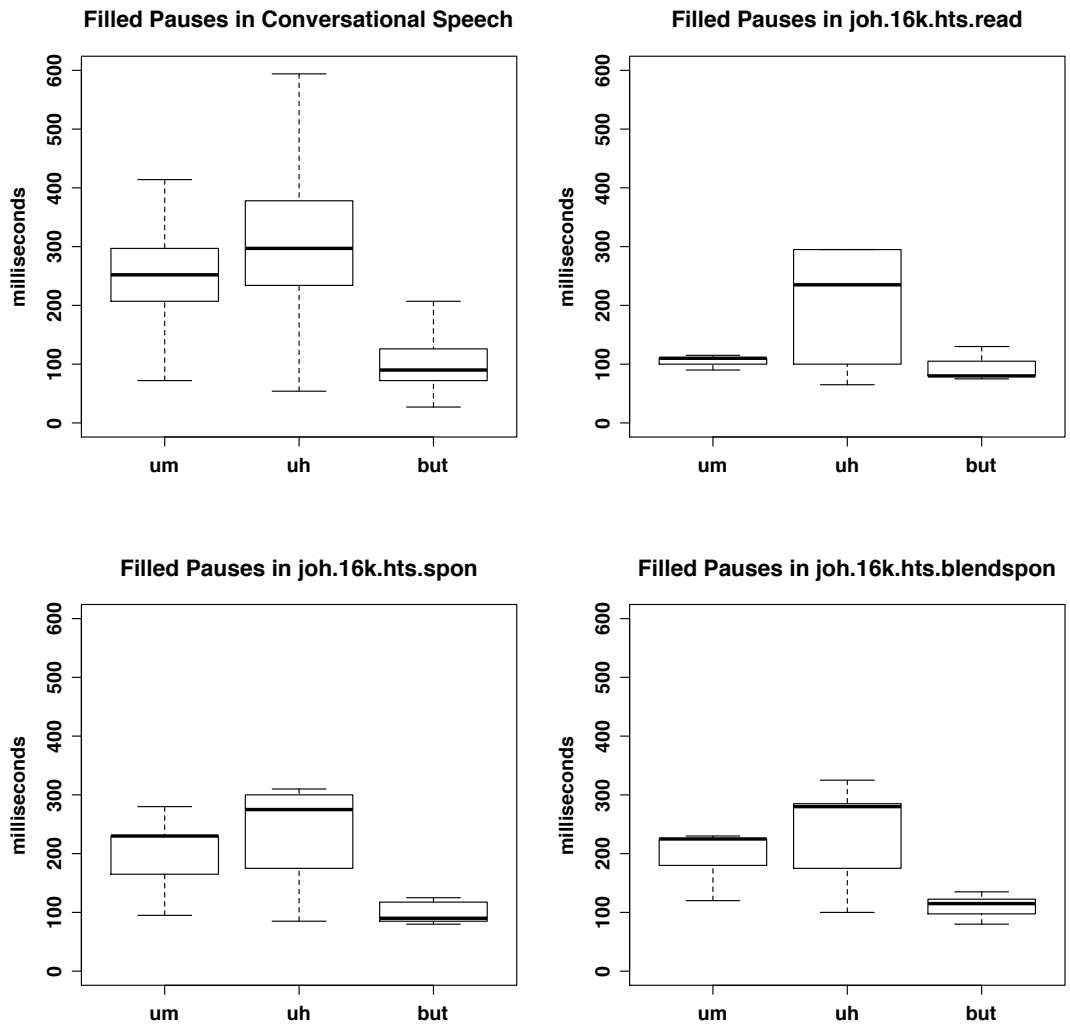

Figure 4.5: Duration of the vowel in the filled pauses (*um* and *uh*), and in the reference word *but*, for natural and synthetic speech. *But* was used as reference because it was represented in the lexicon as having the same vowel quality as the filled pauses, and existed in both the natural and synthetic speech.

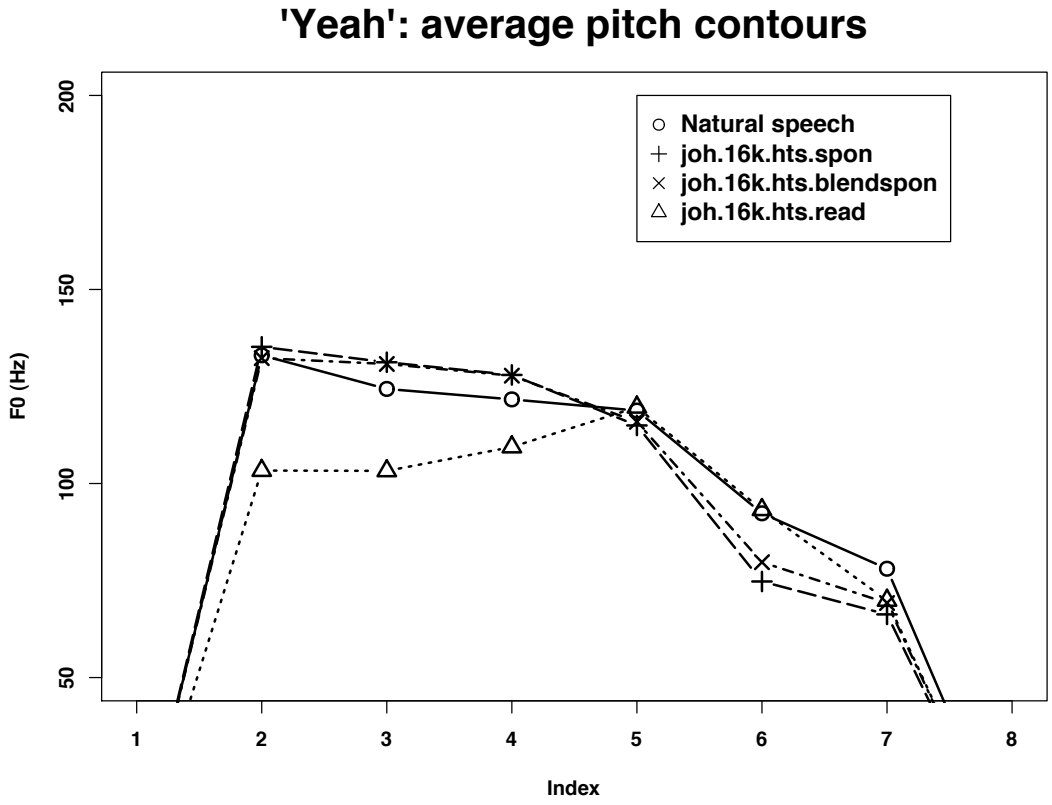

Figure 4.6: Average pitch contours of utterance initial *yeah* in natural and synthetic speech. The $F0$ was measured at every $1/8$ of the total duration for each token.

elling in HMM-based speech synthesis was a joint project between primarily Leonardo Badino and the author of this thesis. L. Badino made predictions of pitch accent and emphasis placements in the training and test sentences, and was primarily responsible for the design and analysis of the perceptual evaluation in Badino et al. (2009). The author designed the context-dependent phonemes that, in addition to the conventional segmental and prosodic categories in neutral read aloud speech, enabled control of emphasis placement and generation in the synthetic speech. The author also built the synthetic voices used in the evaluation in Badino et al. (2009).

4.2.7.1 Emphatic HMM-based Voice

The speech data used for building the HMM-based voice capable of synthesising emphasis was recorded by Strom et al. (2006). The speaker was a male English speaker (Roger). 1132 recorded utterances from the Arctic database (see section 2.3.1) were selected to obtain general phonetic coverage in our voice. These consisted of sentences from fiction that were read aloud in a neutral manner, e.g. *"Author of the danger trail*

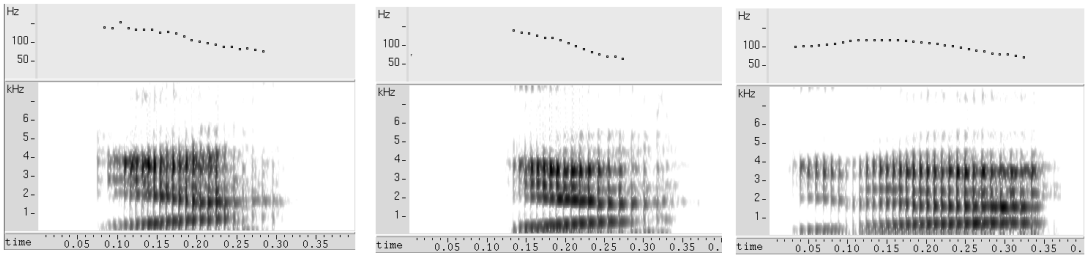

Figure 4.7: A *yeah* in the "same" utterance: natural (left), joh.16k.hts.spon (mid), and joh.16k.hts.read (right). The top pane shows the F0 trajectory, and the bottom pane shows the spectrogram of the different *yeah*-tokens.

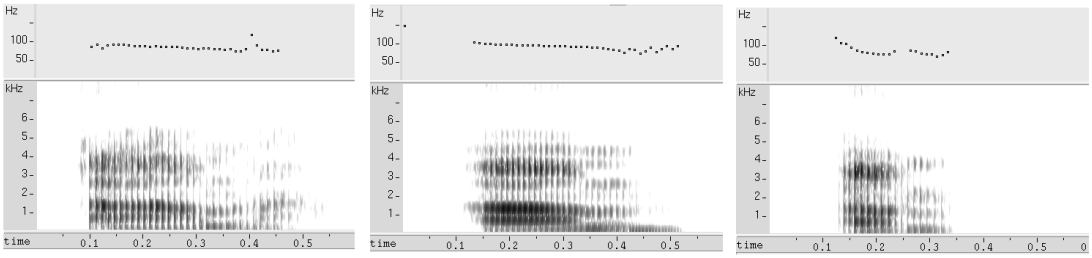

Figure 4.8: The filled pause *um* in the "same" utterance: natural (left), joh.16k.hts-.spon (mid), and joh.16k.hts.read (right). The top pane shows the F0 trajectory, and the bottom pane shows the spectrogram of the different *um*-tokens.

Philip Steels etc." or *"For the twentieth time that evening the two men shook hands."*. The emphasis in Strom et al. (2006) was recorded in stylised carrier sentences of the form:

- It was JAMES who did it.
- No, it was JOHN who did it.
- It was JOHN, not JAMES.

The voice talent was requested to emphasise the upper-cased names. The names were selected to provide diphone coverage of emphasis (Strom et al., 2006). We included these 1683 carrier sentences with the emphasised names together with the Arctic utterances in our voice. In total we had 2815 utterances with approximately 2h of speech data. The voice was built with the standard method outlined in section 4.2.2, but with the conventional contexts in Tokuda et al. (2002) replaced with our alternative context-dependent phonemes described in the next paragraph.

The conventionally used contexts in Tokuda et al. (2002) seemed rather abundant for generating the few prosodic categories that are generally present in conventional synthetic voices. In Badino et al. (2009) we hypothesised that important prosodic categories, e.g. phrase final lengthening and emphasis, could be captured with just a few relevant contexts. In contrast to the counts and positions contexts in Tokuda et al. (2002) we designed a set of contexts within a prosodic window of at most preceding, current and succeeding word:

- {preceding, current, succeeding} phoneme (e.g. b)
- {preceding, current, and succeeding} phoneme types (vowel, plosive, etc.)
- {preceding, current, succeeding} syllable (e.g. b_uh_t_1)
- {preceding, current, succeeding} word[3] (e.g. but)
- current syllable nucleus (e.g. uh)
- pitch accent or emphasis on current syllable nucleus
- pitch accent on {preceding, current, succeeding} syllable and word
- emphasis on {preceding, current, succeeding} phoneme and syllable

The phonemes were clustered based on both articulatory features (plosives, fricatives, etc.) and stress level (for vowels), as in section 4.2.1. The syllable names included the lexical stress (0,1,2). The word level context clustering was only applied to words with frequency above 20 in the training data, which limited the word context to mainly closed class words, and thereby separated function words from content words. A distinction was made between utterance internal and beginning/final silences. We did not include a word level context for emphasis, because we did not want to risk modelling artefacts due to the carrier sentence structure. For example, by including the word context for emphasis, we might have ended up only being able to synthesise emphasis after the word *was*, since that is where the emphasis was in the carrier sentences. The emphasis context on the phoneme level was included instead.

4.2.7.2 Perceptual Evaluation and Discussion

The perceptual evaluation in Badino et al. (2009) was primarily designed and analysed by L. Badino. The part of the evaluation reported in this section was the part where the author was involved in the design and analysis.

[3]The word context was suggested in a discussion with the authors of Raitio et al. (2008) who used it in their voices, but did not mention it in the paper

As part of the perceptual evaluation in Badino et al. (2009) we evaluated whether listeners could identify which word in a synthesised sentence was emphasised. All test sentences are shown in table A.2, in appendix A. Two examples are shown below:

- The more lot came with the HOUSE and the lower the price.
- They tried both soft CONVERSION and hard conversion.

The test sentences were designed by L. Badino and synthesised by the author with the upper-cased word emphasised. The majority of the other open class words were assigned pitch accents and the remaining function words were unaccented.

Thirty-six English native speakers took part in the evaluation. The significance testing was calculated by L. Badino with the binomial test and it showed that the listeners identified the emphasised word in 12 out of the 14 sentences ($p < 0.001$).

The conclusion drawn from the experiment was that we could synthesise emphasis that was significantly more prominent than the pitch accents. Hence, novel prosodic categories can be introduced and controlled in HMM-based speech synthesis via the context-dependent phonemes. But, in the authors' opinions, there was no substantial improvement from using the alternative contexts on the previously modelled prosodic phenomena, e.g. pitch accents and phrase boundaries. Contexts including novel prosodic categories could just as well be added to the default set of contexts. This adding of emphasis contexts to the default set of contexts was later proven to work in Badino (2010), where the current author was not involved.

4.2.7.3 Alternative Contexts for Conversational Speech

In the automatically extracted representations of conversational speech in section 4.2.1 the filled pauses in particular were awkwardly represented. The filled pauses were analysed as pitch accented content words pronounced with the vowel /ʌ/. These properties did not correspond very well to the analyses made by previous researchers in section 2.2.2 or our phonetic analysis in section 3.5.2. The hypothesis was that the specific phonetic properties of the phonemes in filled pauses, and discourse markers, would be better captured by the word-type context used in Badino et al. (2009), because a word context would identify many of these speech phenomena on a token level. For example, the filled pauses would be distinguished from other speech phenomena by being the only words represented as *um* or *uh*. The word context was therefore included in the default set of contexts in section 4.2.1.

	A	B	Equal
Conversational	15%	15%	71%
Naturalness	19%	21%	60%

Table 4.2: Percentage of listeners' preferences of naturalness and conversational speaking style for default (B) and alternative (A) context representations. From a pilot listening test with 8 listeners and 6 sentence pairs. The sentences in the listening test were taken from the held-out conversational material in section 4.2.6 and are shown in table A.3 in appendix A.

We built a voice with the HMM blending method in section 4.2.5, where the word context was added to the contexts in section 4.2.1. The speech data was the same read aloud and conversational speech data as in section 4.2.5. But, no substantial or consistent improvement was perceived on filled pauses or discourse markers (or propositional content) compared to the default contexts. A pilot listening test confirmed that the differences between contexts with and without the frequent word types were at best small (see table 4.2), and in the final perceptual evaluation (in section 5.2) only the default contexts were used. We believe that the amount of discourse markers and filled pauses in limited phrasal contexts in the conversational data (see figure 3.4) was the key to their quality, not their precise representation. This conclusion is supported in the phonetic analysis of the synthetic and natural filled pauses in section 4.2.6 where despite their seemingly odd linguistic representation as pitch accented words with vowel /ʌ/ the synthetic filled pauses have phonetic properties similar to the natural speech.

4.2.8 Summary: HMM-based Voices

Our hypothesis was that by including speech from a spontaneous conversation in HMM-based voices, the voices would convey an impression of a conversational style to listeners. To provide a contrast to these conversational voices we compared them to voices built from conventional "neutral" read aloud sentences.

The pilot experiment in section 4.2.3 attempted to make a conventional read aloud HTS voice exhibit more natural conversational characteristics by adapting it with speech from a spontaneous conversation. This adaptation did not result in a favourable perceptual distinction between the original read aloud voice and the adapted conversational voice. Our conclusion was that this was due to difficulties with adapting to phonetic properties of novel speech phenomena that exists in the adaptation data, but not in

the original voice, in our case the discourse markers and filled pauses. Therefore, we focused on building conversational HMM-based voices with other techniques than adaptation.

In section 4.2.4, we built style-dependent voices from either conventional read aloud sentences or speech from a spontaneous conversation. Two voices were built:

- joh.16k.hts.read: built from the 103min read aloud data in table 3.2
- joh.16k.hts.spon: built from the 75min conversational data in table 3.2.

In general, better phonetic coverage results in better synthetic speech quality (see section 2.3.1). Therefore, we built a voice from all the read aloud and conversational data. We applied a blending technique (see section 4.2.5) that allowed boosting the phonetic coverage compared to the smaller style-dependent voices, while maintaining a distinction between the two speaking styles in the different data sources. This blended voice contained a speaking style parameter to enable switching between speaking styles:

- When we synthesise a "spontaneous" speaking style with the blended voice, we refer to it as joh.16k.hts.blendspon.
- When we synthesise a "read aloud" speaking style with the blended voice, we refer to it as joh.16k.hts.blendread.

The analysis of the phonetic properties in section 4.2.6 showed that both the joh.16k-.hts.spon and the joh.16k.hts.blendspon voices preserved the phonetic properties of frequent conversational speech phenomena. The perceptual evaluation of the style-dependent and blended HMM-based voices is described in section 5.2.

4.3 Unit Selection Voices

The blending and voices described in sections 4.3.3 and 4.3.4 were used for the joint publication Andersson et al. (2010a). The blending method design and building of the blended voice in the original publication were made by the author.

4.3.1 Building CereVoice voices

As described in section 2.3.2 the CereVoice speech synthesis system is based on the conventional unit selection framework. In this section we will describe the voice building procedure with the CereVoice system. The voice building scripts that were used to build the voices in this thesis were developed by CereProc.

The CereVoice speech synthesis system allowed building of 16000Hz and 22050Hz voices. 22050Hz is commercial standard and 16000Hz is used in most referenced work in this thesis. All the speech data was therefore downsampled from 48000Hz to 22050Hz and 16000Hz. Additionally, the default CereVoice voice building methodology slowed down the speech rate by five percent, and applied energy normalisation and companding. The downsampled and processed speech was then parameterized into line spectral frequencies (LSF). Energy, $F0$ and pitch marks were also extracted from the speech samples, to be used in the calculation of concatenation costs. In the voices built for this thesis, this audio pre-processing was generally followed. Any deviations are given for each built voice in 4.3.4.

After the pre-processing, the speech was forced aligned with the method outlined in sections 2.3.1.1 and 3.4.5. The forced alignment gives for each speech sample the phonemic sequence of the corresponding orthographic transcript, including location and duration of utterance internal silent pauses. Each silence delimited speech sample was cut into "spurts". The unit selection target features were extracted for each spurt from the forced aligned phonemic sequence and the transcript. The target features included the current, preceding and succeeding context for each diphone, e.g.:

- {preceding, current, succeeding} phoneme
- {preceding, current, succeeding} syllable
- {preceding, current, succeeding} syllable stress and accent
- position of {preceding, current, succeeding} syllable in word and spurt

Each target feature has a heuristically set weight of how important it is for high quality synthesis compared to the other target features. In the same way, each acoustic feature around the diphone segment boundary is assigned a heuristic weight denoting its importance for allowing a perceptually smooth concatenation ("join") with other speech segments. Additionally, a set of heuristically weighted target features are set based on the general difficulty of concatenating certain phonemes, e.g. vowels are considered more difficult to concatenate than unvoiced fricatives. Each diphone in the recorded speech is then stored together with its associated target and concatenation feature values.

4.3.1.1 Speech Generation

An input sentence is converted to spurt sized phonemic sequences through a set of heuristic rules and look-up in the pronunciation lexicon for the possible pronunciations for each word. The prediction of where to split into spurts is generally based on the punctuation in the input sentence. Then, the target features are extracted for each spurt from the sentence and the derived phonemic sequence.

The database of diphones created as described in section 4.3.1 is then searched with the Viterbi algorithm to find the optimal sequence (lowest combined cost of the concatenation and target feature costs) of available diphone sized speech segments to concatenate into a synthetic utterance.

4.3.2 Pilot: Spontaneous Unit Selection

Initial experiments with utilising spontaneous speech for unit selection revealed that even a small amount of spontaneous speech resulted in perceptually favourable distinctions between read aloud and spontaneous speech synthesis. But including spontaneous speech also resulted in lower naturalness. A qualitative analysis revealed three important factors that contributed to a lower naturalness: segmentation, data sparsity and blending read aloud and spontaneous speech.

We used the CereVoice (Aylett and Pidcock, 2007) unit selection speech synthesiser (described in sections 2.3.2). The voice was built with Heather's (see section 3.2.1) 24min spontaneous conversational speech supplemented by approximately 2h of read aloud phonetic coverage material. The read aloud material was recorded by CereProc. The use of both spontaneous and read aloud speech was done in this pilot mainly to obtain enough speech to make a voice of reasonable quality, but it was motivated from the fact that we cannot control for phonetic coverage in spontaneous speech and blending with read aloud speech would address this problem.

Due to large speech rate differences between the spontaneous and read aloud material, the spontaneous speech was slowed down by 10% and the read speech was speeded up by 5% using SoundTouch's SoundStretch (Parviainen, 2012). The segmentation of the spontaneous speech was done with the method in section 3.4.5. The alignment of the spontaneous speech was far from perfect and often failed in laughing speech, other "extreme" voice qualities and for pronunciation variants that were not represented in our lexicon. The alignment was manually spot-checked and utterances where the alignment clearly failed were removed from the voice build.

During synthesis we utilised CereVoice's genre pruning (see section 2.3.2.1) and biased our selection towards using spontaneous units by pruning out read aloud units if we had a certain number spontaneous units of the correct type. This gave for the utterances in our evaluation on average 69% spontaneous units ranging from 38% up to 93%.

The "spontaneous" voice is henceforth referred to as hea.16k.unit.blend. The read aloud voice used for comparison, henceforth hea.16k.unit.read, was built by CereProc and contained substantially more speech data than the spontaneous voice. Both voices were built with speech from Heather.

4.3.2.1 Perceptual Evaluation

Nineteen held-out utterances from Heather's spontaneous speech were synthesised with the spontaneous (hea.16k.unit.blend) and read aloud (hea.16k.unit.read) voices. These utterances are shown in table A.6 in appendix A. The utterances were presented in pairs to volunteering listeners (both native and non-native English speakers). The listeners were asked to judge which utterance in the pair had *the most spontaneous speech quality* and which had *the best general speech quality, regardless if it sounds spontaneous or not* or if they were equal in any of these aspects. The order of the speech between and within pairs were randomised for each listener. Twenty-two listeners took part in the evaluation.

4.3.2.2 Results

The perceptual judgements have been collapsed over all utterances and are shown in figure 4.9. The significance of the result was tested using the binomial test. The number of times listeners judged the quality as equal ("No preference") was removed before calculating the results.

The results showed that the unit selection voice with spontaneous speech (hea.16k-.unit.blend) was perceived as significantly ($p < 0.001$) more spontaneous than a unit selection voice built with only read aloud speech (hea.16k.unit.read). But the hea.16k-.unit.blend voice also had a lower general speech quality than the hea.16k.unit.blend voice ($p < 0.05$).

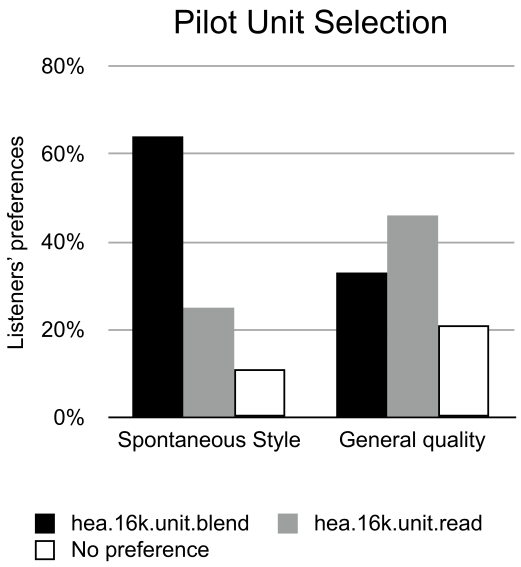

Figure 4.9: Raw data chart. Perceptual judgements of spontaneous and general speech quality in unit selection voices containing spontaneous speech (hea.16k.unit.blend) or just read aloud speech (hea.16k.unit.read). The "No preference" shows the proportion of listeners who did not express any preference between the voices.

4.3.2.3 Diphones and the Perception of Spontaneity

Figure 4.10 shows that utterances with more spontaneous diphones had scores which correlated weakly with the perception of spontaneity. However, the "outlier" in the bottom right corner suggested a modified interpretation of that tendency: it was also important which words contained spontaneous units. The outlier, together with the other three utterances with the highest perceived spontaneity all contained filled pauses that were selected from the spontaneous speech. The fifth and last utterance with a filled pause was only perceived as spontaneous by 50% of the listeners, but it also contained only 58% units from spontaneous speech and was very long.

4.3.2.4 Discussion and Conclusions

Including spontaneous speech in unit selection significantly influenced the perceptual impression of spontaneity in synthetic speech. But we could not compete with the speech quality of the read aloud voice.

Figure 4.10 showed a tendency for unit selection that more spontaneous units increased the perception of spontaneity. But the utterances with 70-80% spontaneous units were perceived as spontaneous in 36-76% of the cases, and we believe that an

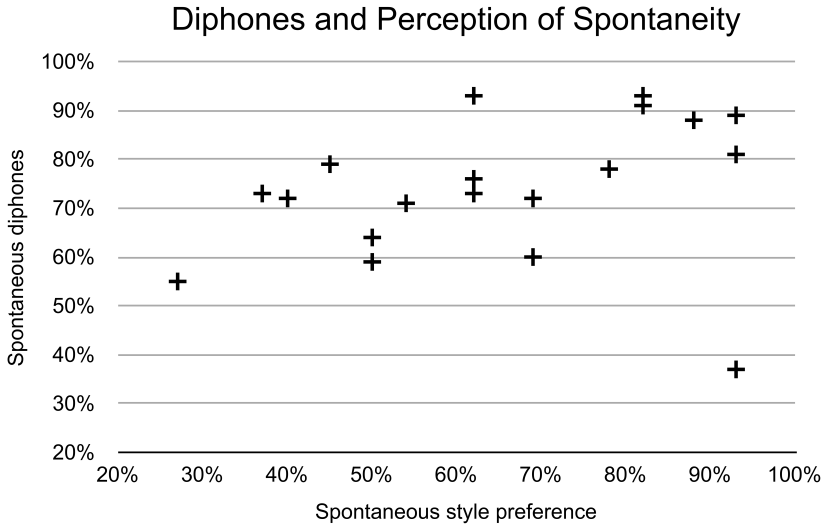

Figure 4.10: Each cross in the figure represents an utterance from the hea.16k.unit-.blend voice in the pilot experiment in section 4.3.2. The cross is placed at the intersection of the percentage of times this utterance was perceived as spontaneous (x-axis) and the percentage of diphones from spontaneous speech in this utterance (y-axis). Spearman's $rho = 0.44$ was non-significant with 95% confidence interval ($p = 0.06$).

important factor of this large span of perceived spontaneity was the blending of spontaneous and read aloud units. Our impression is similar to the conclusions of differences between read aloud and spontaneous speech in Blaauw (1992) in that unstressed syllables seem to be better suited for blending than stressed syllables. But it is also crucial that some units are spontaneous, e.g. filled pauses, which is likely to be the reason for the "outlier" in the bottom right corner of figure 4.10 where "and uh" is from spontaneous units and the rest of the utterance "I met this girl who was Welsh" read aloud units.

4.3.3 Blending Read Aloud and Conversational Speech

In section 2.3.2.1 it was described how CereVoice's genre pruning was used to bias selection towards selecting units from a particular genre, and "backing off" to selecting neutral read aloud units when there was a lack of genre specific units. This genre pruning was applied to the hea.16k.unit.blend voice built from Heather's read aloud and conversational speech described in the pilot experiment in section 4.3.2. The pilot showed some potential for developing conversational speech synthesis with this technique in hea.16k.unit.blend. But, this technique alone did not give a satisfactory result

for Johnny's synthetic speech. The amount of conversational speech from Johnny also meant that the coverage was better, to an extent that when biasing selection towards conversational speech units, hardly any units were selected from the read aloud speech. Although this might sound like a good thing for conversational speech synthesis, it was not. Instead, it often resulted in that more units were selected from less appropriate contexts. Detailed analysis of the current target features and weights could potentially provide minor improvements to speech quality. But, data sparsity is always an issue for unit selection and the target costs are designed to find the best *available* unit. The problem with conversational speech is that if there is no candidate unit with a low target cost the best available alternative is almost certainly not good enough.

Rather than attempting to adjust the target features and weights, the biasing was used to more efficiently achieve a similar effect. For a given input text sentence: if a two-word sequence in the input sentence existed in the conversational speech database then selection was biased towards conversational units, otherwise it was unbiased. This did not guarantee that units were selected from this sequence in the conversational speech database, but it did guarantee that we had suitable candidate units for the given word sequence. The two-word sequence was chosen over a single word to avoid too many genre switches within an utterance. Below are two examples from held-out conversational speech showing the words biased towards conversational speech in bold face:

- **uh it's um a different** character **for me**

- **yeah so it's all up to you guys to make me, yeah** sound good or bad **or whatever**

In the examples above, the text in bold face spans several words, but it is only the two-word sequences, "uh it's", "it's um", "um a", "a different", etc., that were in the conversational data. The longer word sequences were often not in the data, which required bridging the two-word sequences with subword (diphone) units. If the two-word sequence was not in the conversational data, e.g. "different character", then we did not bias selection towards conversational units for the word "character" in this example. In this thesis we refer to this selection bias based on the speech data in the voice as language model bias.

4.3.4 Read Aloud and Conversational Voices

The composition and coverage of Johnny's conversational and read aloud speech data were described in section 3.5, and an overview is shown in table 3.2. The voice used as a baseline in the evaluation in section 5.5 was built by CereProc from 22kHz samples of the read aloud speech, and is henceforth referred to as the joh.22k.unit.read voice in this thesis. To build a synthetic voice with a conversational speaking style we added the conversational speech data in table 3.2 as a genre (see section 2.3.2.1). The segmentation of the conversational speech was described in section 3.4.5, and in the following sections we will describe how the conversational speech was added to the joh.22k.unit.read voice and used to build a voice capable of synthesising conversational characteristics (henceforth referred to as the joh.22k.unit.blend voice in this thesis).

The read aloud speech used in the joh.22k.unit.read voice was downsampled from 48kHz to 22kHz, power normalised and companded. The conversational speech was therefore processed in the same way.

One of the more noticeable differences between the conversational and read aloud data was the speaking rate (see section 3.5.2.3). As mentioned in section 4.3.1 the read aloud speech was slowed down five percent, and to facilitate blending of conversational and read aloud material in synthetic utterances the conversational speech was slowed down with ten percent. There was also more variation in speaking rate in the conversational than in the read aloud speech (see section 3.5.2.3). But as a starting point it was considered sufficient to reduce the conversational speaking rate with a constant value. The alternative to speed up the read aloud speech was rejected on the basis that:

- it was considered unnecessary to tamper with CereVoice's standard voice building method and risk a negative impact on quality of the read aloud voice
- the conversational characteristics that we focused on was discourse markers and filled pauses, not speaking rate.

4.3.4.1 Blending or Not

In the joh.22k.unit.blend voice the conversational speech was added to the read aloud data with the blending technique described in section 4.3.3. But, perhaps there is no need for a blending technique, perhaps we could just combine the conversational and read aloud speech to achieve our aim of natural and conversational speech synthesis?

Two voices were built with the combined conversational and read aloud data in table 3.2. The voice, henceforth referred to as joh.16k.unit.baseline, was built with the default method in section 4.3.1: all the conversational and read aloud data were pooled, downsampled to 16kHz, power normalised, companded and slowed down five percent. The other voice, henceforth joh.16k.unit.blend, was built using the same genre blending method as the joh.22k.unit.blend voice (the method is described in section 4.3.3). The speech data for the joh.16k.unit.blend was processed the same way as the joh.16k-.unit.baseline voice except that the conversational speech was slowed down by ten percent instead of five. Additionally, the joh.16k.unit.blend voice utilised the segmentation technique described in section 3.4.5, whereas the joh.16k.unit.baseline voice was force aligned with CereVoice's standard method described in section 2.3.1.1. In the joh.16k.unit.baseline voice all the data were pooled also for training the forced alignment models, which could be a better method given that there is more training data and all the data that subsequently will be forced aligned is used for training the models.

The perceptual evaluation of the joh.16k.unit.baseline and joh.16k.unit.blend voices is described in section 5.4.

4.3.5 Filler Prediction

The filler prediction described in this section was made by Kallirroi Georgila for the joint publication Andersson et al. (2010a). This section is included as part of this thesis to describe how the test sentences were generated for the evaluation in section 5.5.

If speech synthesis capable of synthesising a conversational speaking style was used for a believable character, a symbolic representation of the speech (e.g. a word sequence) would be passed to the synthesiser to convert to a speech signal. Some of that content, e.g. discourse markers and filled pauses, might be sensible to predict both with respect to when they should be used, but also with respect to the specific language use of the character, i.e. the specific speech data in the voice. For the work in Andersson et al. (2010a) we focused on the latter part to show what the voice was capable of saying. For example, from the speakers in chapter 3, Roger used *yeah* a lot less than Johnny and Heather, and to preserve Roger's character and personal preference of expressions his voice should probably say *yes* or *right* instead.

To predict what Johnny's voice could say, Kallirroi Georgila implemented a "filler" (discourse marker or filled pauses) prediction algorithm. The filler prediction was

based on the speech data that was included in the synthetic voice and was a prediction of what the voice was likely to be able to synthesise. Therefore the prediction of filler sequences was only evaluated with respect to how they sounded in the synthetic speech, and not with respect to e.g. text based precision and recall.

The prediction algorithm was described in detail in Andersson et al. (2010a) and an overview is given in figure 4.11 and here:

- calculate n-gram probabilities for the 2120 conversational utterances in the voice
- when given a test sentence without fillers, e.g. *<s> it's a miracle </s>*
- look up in the conversational speech data existing fillers that followed any of the words in the input sentence, e.g. *it's uh*, *it's like*, *a um*, *a you know*, etc.
- repeat the previous step to generate potential sequences of fillers, e.g. *it's um uh*, *miracle um yeah*.
- use Viterbi decoding to find the sequence of propositional content and fillers with the highest probability, e.g. *so it's a miracle um yeah*

The CereVoice unit selection engine does the Viterbi search for silence delimited speech sequences. The filler prediction algorithm did not include predictions of silence around the filler sequences. By inserting silences, we would avoid some of the otherwise required diphone joins in an utterance. Since the joins are one of the most critical factors when doing unit selection, we designed rules for when to insert silences around the predicted fillers. The silence insertions were designed by the author and worked as follows:

- If there was a genre switch, before or after the filler sequence, insert a silent pause before or after respectively. This was done because a genre switch suggested that we did not have that sequence in the data.
- Insert a pause after a sentence initial filler, unless there was an inserted "filler word filler" sequence (e.g. *uh it's um*). A predicted "filler word filler" sequence suggested that the sequence existed in the voice data, and it could therefore be selected as a contiguous sequence.
- Insert a silent pause before an utterance final filler, to increase chances of getting phrase final units for the word preceding the filler.

4.3.6 Properties of the Synthetic Speech

The unit selection framework of selecting units directly from the recorded speech, caused very few units from discourse markers and filled pauses to be selected from other words than discourse markers and filled pauses when the genre blending was applied. To test the general feasibility of our approach, the set of 169 held-out utterances from section 4.2.6 was synthesised with the blended joh.22k.unit.blend voice. These held-out sentences are a good representation of what we would like to be able to synthesise with conversational speech synthesis. When genre blending was turned on, 72% (1769/2461) of the words were biased towards selecting units from the conversational speech. Table 4.3 shows how many diphone units were selected for these 169 sentences when genre blending was on or off. The conclusions drawn are, a) that many words and diphones are already in the recorded conversation, and b) that many are not, and blending is therefore necessary, in order to use unit selection. To synthesise, e.g. discourse markers and filled pauses, with natural phonetic properties we can select these directly from a spontaneous conversation, but in order to wrap these around unrecorded propositional content we also need standard subword unit selection.

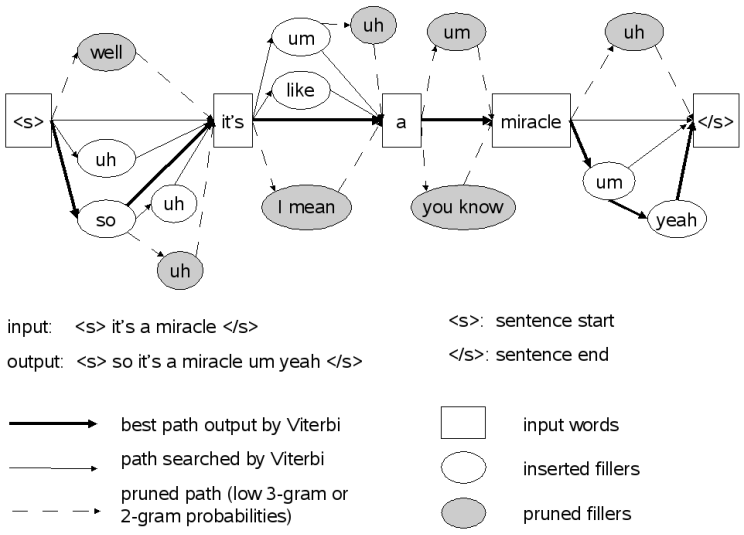

Figure 4.11: Example of filler prediction. (The figure was made by Kallirroi Georgila for Andersson et al., 2010a). The figure shows an input sentence, *<s> it's a miracle </s>*, in the square boxes. The possible transitions are shown with arrows, and possible filler words are shown in the ovals. The Viterbi algorithm was used to decode the most likely fillers to insert in the sentence. In this example the output sentence becomes: *so it's a miracle um yeah.*

	Conversational Diphones
Genre Blending "on"	62.2%
Genre Blending "off"	7.5%

Table 4.3: Percentage of diphone units selected from the conversational data when genre blending was "on" or "off". The test sentences consisted of 169 held out utterances from the conversation with Johnny. The total number of diphones in the test sentences was 7521.

4.3.7 Summary: Unit Selection Voices

The aim of this thesis is to build voices that can synthesise high quality conversational style speech. Our approach to achieve that goal was to utilise blending of conversational and read aloud speech. We utilised speech from a spontaneous conversation to synthesise conversational speech phenomena with natural phonetic properties. But, to compensate for the gaps in phonetic coverage in the spontaneous speech data we augmented it with conventional read aloud sentences. This was done because, as the analysis in figure 3.4 showed, it is not feasible to achieve phonetic coverage by recording spontaneous speech.

The pilot experiment in section 4.3.2 showed that by augmenting the conventional read aloud speech with conversational speech in synthetic voices we can convey a more realistic impression of a conversational style. The challenge that we addressed was to find a better trade-off between when we can and should select units from conversational speech, and when it is better to "back-off" to read aloud units. Therefore we focused on developing the blending of conversational and read aloud speech data.

The developed blending method and the voices that we evaluated in chapter 5 are summarised in this paragraph. The unit selection voices were built with either the standard unit selection method or with our developed blending technique described in section 4.3.3. An overview of the conversational and read aloud speech data used in the voices is shown in table 3.2. In total four voices were built:

- joh.22k.unit.read: a standard unit selection voice built from 103min of neutrally read aloud sentences. The voice is described in section 4.3.4.

- joh.16k.unit.baseline: a voice built with the standard unit selection method, but containing both the 75min conversational speech and the 103min read aloud speech. The voice is described in section 4.3.4.1.

- joh.22k.unit.blend and joh.16k.unit.blend: two voices built with the developed blending method. The two voices were built from the same speech data, but the audio was downsampled to 22kHz in the joh.22k.unit.blend voice and 16kHz in the joh.16k.unit.blend voice. The building of the voices are described in section 4.3.4. All the conversational and read aloud data in table 3.2 were included in these two voices, in total 178min. The blending method used when building the voices included the modified forced alignment of the conversational data described in section 3.4.5, the speaking rate adjustment described in section 4.3.4, and the language model bias to select units from either the conversational or read aloud data described in section 4.3.3.

The analysis of the synthetic speech showed that the majority of units in the joh.22k-.unit.blend voice were selected from the conversational data, when synthesising in-domain material. In the perceptual evaluations in chapter 5, the blended voices were contrasted with the voices built using the standard unit selection method.

4.4 Conclusion

The currently dominating speech synthesis frameworks, unit selection and HMM-based speech synthesis, build synthetic voices by modelling the phonetic properties in recordings of natural speech. This creates high quality synthetic voices that sound like the speech and the speaker of the original recording. Therefore, by building unit selection and HMM-based synthetic voices with speech from a spontaneous conversation the voice would exhibit phonetic properties similar to natural conversational speech.

In chapter 3 we described how speech from a spontaneous conversation was recorded, transcribed and segmented. The goal was to obtain conversational speech data that would allow building of high quality unit selection and HMM-based voices. As a result we obtained a data set of 2120 conversational utterances (75min of speech). This data set was rich in conversational speech phenomena, in particular discourse markers and filled pauses (see section 3.5.1). The problem is that it is not feasible to achieve general phonetic coverage by recording only spontaneous speech (see figure 3.4). If we cannot achieve phonetic coverage it will be difficult with current techniques to synthesise high quality utterances that are not pre-recorded.

Our approach to address this coverage issue in spontaneous speech resources was to blend the conversational speech data with conventional recordings of read aloud sen-

tences. The read aloud sentences were recorded to provide general phonetic coverage. In total we had 75min conversational data and 103min read aloud data (see table 3.2). By blending these two data sources, we aimed to create synthetic voices that could preserve the natural phonetic properties of conversational speech phenomena while boosting the phonetic coverage with read aloud speech to maintain consistently high quality synthesis.

The unit selection blending was described in section 4.3.3. The unit selection blending was designed to select units from the conversational speech data only when we could be fairly certain that appropriate units existed in the conversational data. The decision to bias selection towards conversational units was made through a language model. In summary, when words in an input sentence existed in the recorded conversational speech data, then the units were selected from the conversational data. Otherwise the units were selected from the read aloud data, to reduce the risk of introducing acoustic artefacts in the synthetic speech. The analysis of the blending in section 4.3.6 shows the feasibility of the method: when synthesising in-domain conversational material the majority (62.2%) of the units were selected from the conversational data. The question is, what impression is conveyed to the listeners:

- Does the blending result in natural-sounding speech, or is it obvious to listeners that spontaneous and read aloud speech are spliced together in the same utterance?

- Is the proportion of units selected from conversational data enough to convey a general impression of a conversational style to listeners?

The HMM-based blending was described in section 4.2.5. HMM-based speech synthesis estimates statistical distributions of phonetic properties from recordings of speech data. The larger amount of data obtained when combining the conversational and read aloud speech was intended to result in more reliably estimated phonetic properties, and hence better quality synthetic speech. The blending was enabled by adding a speaking style context, spontaneous or read aloud, to the context-dependent phonemes. This context was then available during training of the voice to preserve distinguishing phonetic properties between the two speaking styles in the source data (see section 3.5). The blending technique was selected because it had been proven successful in preserving distinctions between other speaking styles: joyful, sad or rough (Yamagishi et al., 2005). When generating speech, the speaking style context was set on the utterance level to either read aloud or spontaneous, to "bias" the generation of synthetic speech

towards the phonetic properties of either read aloud or spontaneous speech. The phonetic analysis in section 4.2.6 of the synthetic speech showed that the blended voice preserved phonetic properties of discourse markers and filled pauses as well as a voice built from only the conversational speech data. But, there were also tendencies that blending smoothed out the differences, e.g. in speaking rate, between the conversational and read aloud data. Therefore, we will investigate the impression conveyed to listeners:

- Do style-dependent HMM-based voices, built from either conversational or read aloud speech, convey a distinction between speaking styles?
- Does the blending in a HMM-based voice preserve a distinction between read aloud and conversational speaking styles?

In chapter 5 we will describe the perceptual evaluations made on the unit selection and HMM-based voices.

Chapter 5

Perceptual Experiments

This chapter describes the perceptual experiments that were made on the synthetic voices built in chapter 4. The general hypothesis tested was: does the inclusion of conversational speech in a synthetic voice add conversational characteristics to the synthetic voice without a negative impact on naturalness? The specific hypothesis tested with each experiment will be stated in that respective section.

The large scale speech synthesis evaluations in the Blizzard Challenge (see section 2.3) make use of both native and non-native listeners when evaluating the naturalness of synthetic speech, and we have found no literature that suggests that non-native speakers of English make substantially different judgements from native speakers when evaluating naturalness and conversational style. Therefore, we used both native and non-native English speakers as participants in our evaluations.

The experiments presented in this chapter did not compare unit selection voices to HMM-based voices. The experiments focused on investigating the impact of different types of data and our developed blending techniques only within the two speech synthesis frameworks.

5.1 Independent Contribution by the Author: Experimental Design and Analysis

The results from the experiments in sections 5.2 and 5.5 have been published in Andersson et al. (2010a), Andersson et al. (2010b) and Andersson et al. (2012). Both these experiments were designed and analysed by the current author. The decision to use predicted sentences rather than held-out material in the experiment in section 5.5 was made together with Kallirroi Georgila.

The experiments in sections 5.4 and 5.6 have at the time of writing not been reported elsewhere. These experiments were also designed and analysed by the author.

5.2 Evaluating Naturalness and Conversational Style of the HMM-based Voices

In section 2.3.3.6 we claimed that the challenge for HMM-based speech synthesis is to build voices that convey a general impression of a conversational quality or style. The background chapter 2 and the analysis of conversational and read aloud speech in chapter 3 suggested that a key problem for synthesising conversational style speech lie in appropriate synthesis of frequent conversational speech phenomena: discourse markers and filled pauses. The evaluation of the HMM-based voices was therefore designed to test to what extent voices built from conversational speech data; rich in discourse markers and filled pauses, conveyed an impression to the listeners of exhibiting natural conversational characteristics, i.e. having a conversational style. This was evaluated by contrasting voices and utterances where there was a hypothesised difference in conversational style.

The synthetic voices used in the evaluation: joh.16k.hts.spon, joh.16k.hts.read, joh.16k.hts.blendspon and joh.16k.hts.blendread, were described in section 4.2. The speech data used for the different voices was described in chapter 3. The speech data table from chapter 3 is reprinted in this chapter as table 5.1. The joh.16k.hts.spon voice was built from only the 75min conversational data in table 5.1 and the joh.16k.hts.read voice was built from only the 103min read aloud data. These two voices were built with the conventional HTS system (see section 4.2.4). The third voice, which we refer to as joh.16k.hts.blendspon and joh.16k.hts.blendread, is *one* voice. In this voice a blending technique was applied to allow building higher quality voices by combining speech data with different phonetic properties. An important aspect of the blending was that the speaking styles in the source data: conversational or read aloud, could be preserved in the synthetic speech. This voice was built from both the 75min conversational data and the 103min read aloud data. When synthesising the test sentences with the blended voice one of the speaking styles was selected by setting the speaking style context to spontaneous conversational or read aloud (see section 4.2.5). These utterances where the speaking style was set to either conversational or read aloud are referred to as being from the joh.16k.hts.blendspon or joh.16k.hts.blendread voice. The phonetic analysis of the synthetic speech in section 4.2.6 confirmed that the joh.16k.hts.blendspon voice

	Conversation	Read Aloud
utterances	2120	2717
word tokens	19841	22363
word types	2200	5026
syllable tokens	24657	30902
phone tokens	58332	75856
diphone types	1769	2483
quinphone types	37654	58867
total duration (incl. silence)	89min	106min
total duration (excl. silence)	75min	103min

Table 5.1: Overview of Johnny's conversational and read aloud data. The table is a reprint of table 3.2. The duration shows the amount of phonetic material, including or excluding utterance internal silent pauses. The diphone types include silences and lexical stress on vowels. The quinphone types include silences, but not lexical stress.

preserved phonetic properties of the discourse markers and filled pauses as well as the joh.16k.hts.spon voice did. A summary of the techniques and speech data used for the different voices is shown in table 5.2.

A perceptual experiment was designed to test these three voices' ability to synthesise natural sounding conversational characteristics. The experimental design and selection of test sentences are described in section 5.2.1. The experiment tested two hypotheses:

I) A voice built with conversational speech (joh.16k.hts.spon) is more conversational and more natural than a conventional voice (joh.16k.hts.read) when the synthetic utterances contain discourse markers and filled pauses. The reason being the differences in phonetic content of the conversational and read aloud data used to build these voices, where the discourse markers and filled pauses have a high frequency in the conversational data, but are nearly absent from the read aloud data (see section 3.5).

II) Utterances with appropriately synthesised discourse markers and filled pauses (joh.16k.hts.blendspon) are perceived as more conversational, but not less natural, than utterances without discourse markers and filled pauses (joh.16k.hts-.blendread) when we utilise blending. The reason being that blending allows the

Voice	System	Sampling rate	Blending	Read data	Conv. data	Total data	Experiment
joh.16k.hts.spon	HTS	16kHz	no	-	75min	75min	sec. 5.2
joh.16k.hts.read	HTS	16kHz	no	103min	-	103min	sec. 5.2
joh.16k.hts.blend[spon|read]	HTS	16kHz	yes	103min	75min	178min	sec. 5.2
joh.16k.unit.baseline	CereVoice	16kHz	no	103min	75min	178min	sec. 5.4
joh.16k.unit.blend	CereVoice	16kHz	yes	103min	75min	178min	sec. 5.4
joh.22k.unit.blend	CereVoice	22kHz	yes	103min	75min	178min	sec. 5.5 & 5.6
joh.22k.unit.read	CereVoice	22kHz	no	103min	-	103min	sec. 5.5 & 5.6

Table 5.2: Summary of techniques and speech data used for the different voices. All these voices were built with Johnny's speech. The amount of speech data is given in minutes; 103min read aloud speech and 75min conversational speech. More details of the speech data and the building of the different voices can be found in chapters 3 and 4.

synthetic voice to combine high quality synthesis of speech phenomena typical for conversational speech, as well as the arbitrary propositional content that conventional voices synthesise so well.

Sections 5.2.1, 5.2.2, 5.2.2.1 and 5.2.2.2 contain material that were originally published as collaborative work in Andersson et al. (2010b) and Andersson et al. (2012). The evaluation design and analysis of the results in the original publications were all made by the author of this thesis.

5.2.1 Evaluation Design

The test sentences for the listening test were randomly selected from the synthetic utterances in section 4.2.6, but with restrictions on the syntactic and semantic content, so that they contained at least two discourse markers or filled pauses and were between 5-15 words long in total, e.g. *oh yeah you don't want that to happen*. These sentences were synthesised with the joh.16k.hts.spon, joh.16k.hts.read and joh.16k-.hts.blendspon voices. To test hypothesis I, stated in section 5.2, the utterances from the joh.16k.hts.spon were compared to the utterances from the joh.16k.hts.read. To test hypothesis II, the discourse markers, filled pauses and disfluencies were removed to obtain more conventional sentences, e.g. *you don't want that to happen*. These sentences were synthesised with the joh.16k.hts.blendread voice, and compared to the joh.16k.hts.blendspon utterances. Examples of the compared utterance pairs are shown in table 5.3. All test sentences are shown in appendix A, tables A.4 and A.5.

To avoid a scenario where it was obvious from the text alone that the discourse

markers and filled pauses had been removed from one of the utterances, the compared pairs always contained different utterances, and hence differing lexical content. To exemplify: if we had compared A) *so let's see, but um, yeah, nothing exciting*, to B) *let's see, but nothing exciting*, listeners could easily identify that one utterance had the same lexical content as the other plus/minus a few conversational markers *yeah*, *um*, *oh*, etc. Whereas when we compared the utterances A) *right, oh you have to to transcribe all this*, to B) *let's see, but nothing exciting*, the large differences in content would make it more difficult to identify that a few words had been removed, and hence make it easier to evaluate speaking style and not text content.

Naturalness is conventionally used in speech synthesis to evaluate speech quality, but evaluating a conversational style has been less explored. We suspect that when listeners are asked to judge the quality of synthetic speech they do so in a quite general way by judging based on the most prominent difference between utterances, rather than the specific feature they have been asked to judge, in effect evaluating which utterance sounds "best". To investigate this issue further, the listeners were divided into two groups where each group was requested to evaluate one criteria each: naturalness or conversational style. One group was requested to evaluate: *"Which utterance sounds more like natural speech?"*. The other group was requested to evaluate *"Which utterance has a more conversational speaking style?"*. The participants who were asked about the conversational style were also explicitly requested to disregard the speech quality: *"Please try and disregard the speech quality, and focus on the speaking style."*. This extra request was intended to make the listeners aware that it was not necessarily the most prominent quality difference between the utterances that we wanted them to

Voice	Text	Pair No.
joh.16k.hts.spon	"right, oh you have to to transcribe all this"	5 in
joh.16k.hts.read	"so let's see, but um, yeah, nothing exciting"	fig. 5.3
joh.16k.hts.blendspon	"right, oh you have to to transcribe all this"	5 in
joh.16k.hts.blendread	"let's see, but nothing exciting"	fig. 5.4
joh.16k.hts.spon	"um, but even that like, I can give a shit less, you know what I mean"	11 in
joh.16k.hts.read	"oh yeah you don't want that to happen"	fig. 5.3
joh.16k.hts.blendspon	"um, but even that like, I can give a shit less, you know what I mean"	11 in
joh.16k.hts.blendread	"you don't want that to happen"	fig. 5.4

Table 5.3: Examples of test sentence pairs for the evaluation in section 5.2. Commas indicate where utterance internal silences were located. All the pairs in the evaluation are shown in appendix A, tables A.4 and A.5.

judge. Thirty-two participants, where the majority were native English speakers, were paid to take part in the evaluation.

5.2.2 Results

The results of the evaluation are summarised in figures 5.1 and 5.2. The significance was tested with the binomial test. joh.16k.hts.spon was perceived as significantly ($p < 0.05$) more natural and more conversational than the joh.16k.hts.read. This supports hypothesis I in section 5.2. However, the joh.16k.hts.blendread was perceived as significantly ($p < 0.05$) more natural than the joh.16k.hts.blendspon. Additionally, the joh.16k.hts.blendspon utterances were not perceived as significantly ($p = 0.25$) more conversational than the joh.16k.hts.blendread utterances. In sections 5.2.2.1 and 5.2.2.2 we will further analyse these results for *naturalness* and *conversational style* to investigate why only hypothesis I (see section 5.2) was supported.

5.2.2.1 Naturalness

Figures 5.3 and 5.4 show the participants' perceived naturalness for individual utterances. By listening to the utterances, factors that were likely contributors to the differences between utterances in figures 5.3 and 5.4 were identified.

Some factors were easily identified. In utterances 10 and 11 in figure 5.4 there were prominent local pitch errors in the joh.16k.hts.blendspon utterances. In the joh.16k.hts-.blendspon utterances 5 and 15, the first had a function word repetition that was too prominent, and in the latter the utterance prosody was not natural. To some extent these errors are caused by the underspecified analysis and representation of segmental and prosodic properties in conversational speech in our synthetic voices. But, the prominent word repetition in utterance 5 was prominent also in the original natural utterance, and the participants' judgements were perhaps negatively influenced by the presence of an audible disfluency, a factor that we also discuss in the next paragraph.

A general factor in the perceived naturalness was that a) the filled pauses and discourse markers (in particular *yeah*) sounded bad with the joh.16k.hts.read and b) when the discourse markers, filled pauses and disfluencies were removed in the sentences synthesised with the joh.16k.hts.blendread voice, it made them sound substantially better. The removal of discourse markers, filled pauses and disfluencies also made many utterances more grammatical and more fluent than the original conversational utterances, e.g. utterance pairs 2, 3, 4, 9, and 14 (see table A.5), which may have

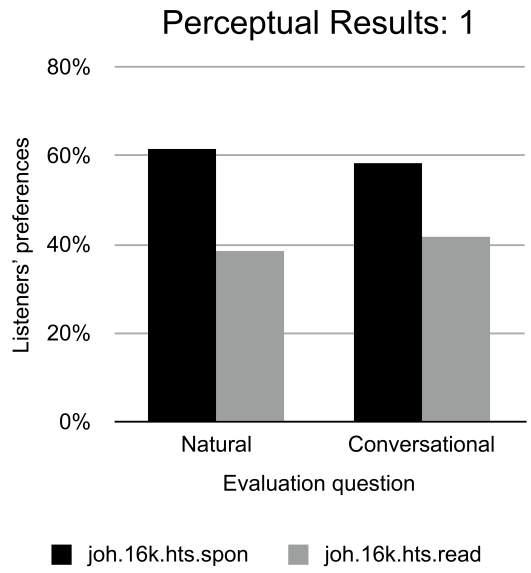

Figure 5.1: The bars show the percentages of the participants' preferences for naturalness and conversational style when comparing the joh.16k.hts.spon to the joh.16k.hts-.read.

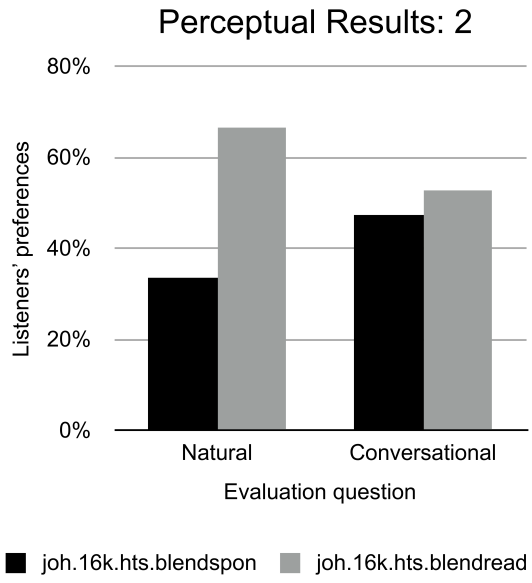

Figure 5.2: The bars show the percentages of the participants' preferences for naturalness and conversational style when comparing utterances with conversational characteristics (joh.16k.hts.blendspon) to more fluent utterances (joh.16k.hts.blendread).

contributed to making the perceived differences in naturalness larger than they were.

5.2.2.2 Conversational Speaking Style

The perceptual evaluation was designed to see if we could evaluate conversational speaking style separately from naturalness. The questions about naturalness and speaking style were therefore asked to separate groups of participants. There is no need for a conversational discourse in order for listeners to identify which utterance has a more conversational style, see sections 2.2 or 2.3.3.6 for a motivation of this.

The speaking style results, shown in figure 5.1, were significantly in favour of the joh.16k.hts.spon, but so were the results for naturalness, and the correlation between them was significant (Spearman's $\rho = 0.72, p < 0.001$). Our interpretation was that the difference between the voices in figure 5.1 was a difference in naturalness rather than speaking style.

There was no significant difference in the perceived speaking style for the joh.16k-.hts.blendspon and joh.16k.hts.blendread voices. However, the correlation between the two groups' perceptions of naturalness and speaking style was very strong (Spearman's $\rho = 0.86, p < 0.001$), visualised in figure 5.5. This indicates that for an utterance to be perceived as having a conversational speaking style, it also needs to be perceived as fairly natural. Even without discourse markers and filled pauses, the test sentences contained other conversational, or casual, characteristics, e.g. *...I could give a shit less...*, *...cool* or *...kind of a freak*, which contributed to making the evaluation of speaking style more difficult.

Figure 5.6 shows individual participants' judgements of conversational speaking style, which shows that there were at least two different interpretations of speaking style, where participants *a-d* have interpreted speaking style differently than participants *l-p*. In contrast, only one participant perceived the joh.16k.hts.blendspon utterances as more natural than the joh.16k.hts.blendread utterances.

5.2.3 Conclusion

The evaluation in section 5.2 was designed to test the HMM-based voices ability to convey an impression of a conversational style to listeners. The evaluation therefore contrasted speech with and without potentially conversational characteristics, as outlined in section 5.2 and summarised here:

I) A voice built with only conversational speech (joh.16k.hts.spon) was contrasted

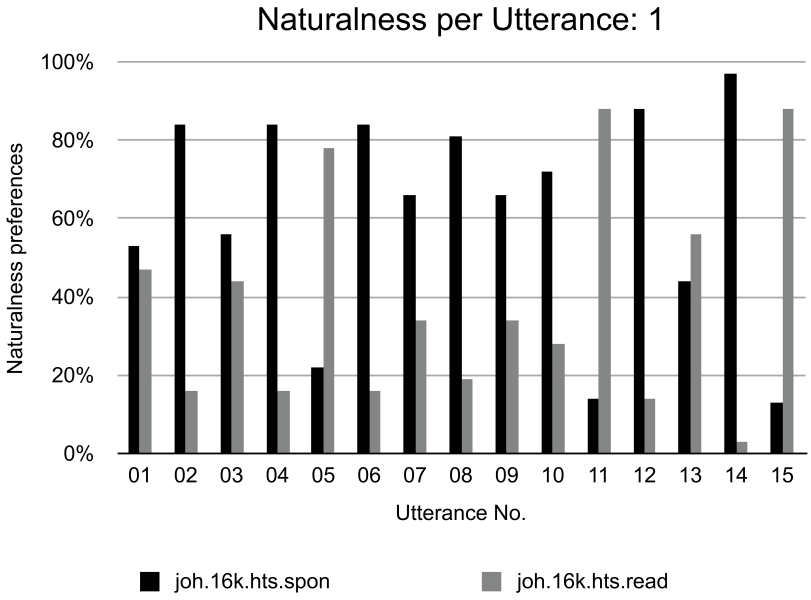

Figure 5.3: Participants' perception of naturalness for individual utterances when comparing sentences with discourse markers and filled pauses synthesised with the joh.16k.hts.spon or joh.16k.hts.read voices.

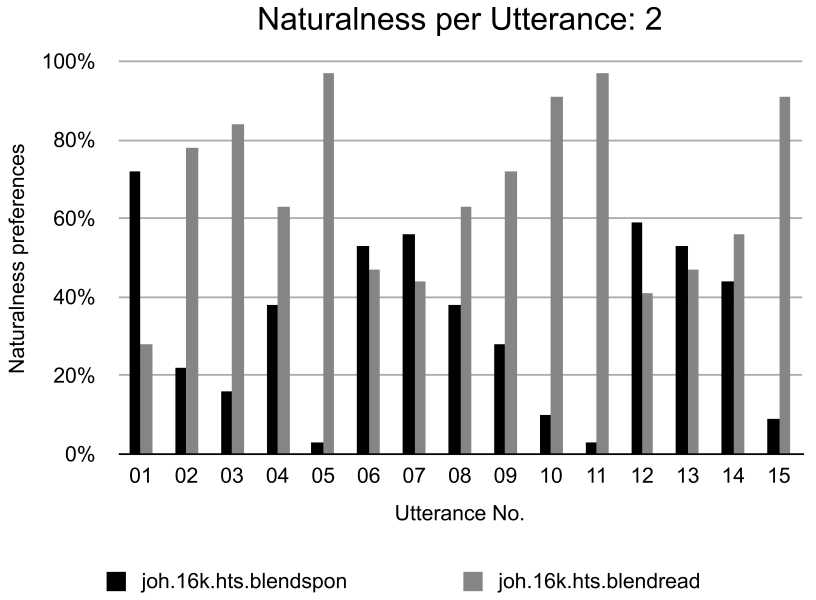

Figure 5.4: Participants' perception of naturalness for individual utterances when comparing sentences synthesised with the blended voice. The joh.16k.hts.blendspon bar shows preference for utterances with conversational characteristics, and the joh.16k-.hts.blendread bar shows preference for more fluent utterances.

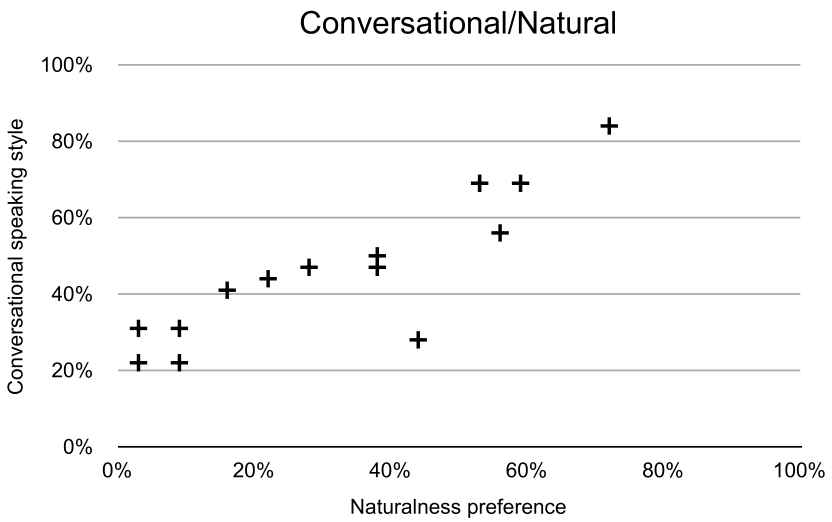

Figure 5.5: Plot of the participants' preferences of naturalness and conversational speaking style for the joh.16k.hts.blendspon voice. Spearman's rho showed significant ($p < 0.001$) correlation of 0.86.

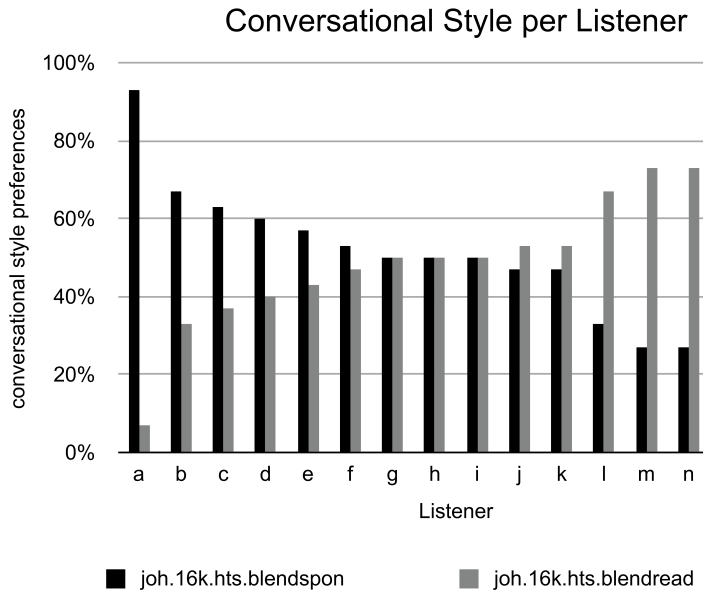

Figure 5.6: Individual participants' perception of conversational speaking style when comparing sentences synthesised with the blended voice. The listeners are ordered from left to right in the figure based on their judgements of how conversational the joh.16k.hts.blendspon voice was.

with a voice built with only read aloud speech (joh.16k.hts.read). The speech data and techniques used for building the HMM-based synthetic voices were described in chapter 4 and a summary is shown in table 5.2.

II) Utterances that contained frequent conversational speech phenomena; discourse markers and filled pauses, were contrasted with utterances that did not contain these conversational characteristics. Examples of the compared sentences are shown in table 5.3. The utterances with discourse markers and filled pauses were synthesised with the joh.16k.hts.blendspon voice, and the utterances without these speech phenomena were synthesised with the joh.16k.hts.blendread voice. A summary of the voices in the evaluation is shown in table 5.2.

Two separate groups of listeners were requested to judge either the naturalness or the conversational style of the synthetic utterances. The two groups judgements of these two criteria were strongly correlated (see section 5.2.2.2), which suggested that they measured similar aspects of the utterances. This supported our suspicion in section 5.2.1 that listeners often make a judgement based on the most prominent difference, rather than the particular criteria they have been requested to evaluate. Therefore, we cannot conclude with absolute certainty whether the difference between the voices was in naturalness or conversational style. However, we can conclude that there was a difference in some quality aspect which, given the nature of our voices and test sentences, was due to the different voices ability to synthesise conversational speech phenomena.

The results of the perceptual evaluation in section 5.2.2 showed that the voice built with only conversational speech (joh.16k.hts.spon) was perceived by listeners as sounding more natural than the conventional voice (joh.16k.hts.read). Our interpretation was that this result was due to the joh.16k.hts.spon voice's better ability to synthesise propositional content wrapped in discourse markers and filled pauses. However, the joh.16k.hts.blendspon utterances were perceived as less natural than the joh.16k-.hts.blendread utterances, which showed that we could not synthesise these conversational style utterances with as high quality as we could synthesise more conventional "fluent" sentences.

The results shown in figures 5.1 and 5.2 suggested that the joh.16k.hts.spon had a more conversational style than the joh.16k.hts.read, and that the blended voice did not preserve a distinction between the conversational and read aloud speaking styles in the source data. However, supported by the qualitative analysis in section 5.2.2.1 and the

strong correlation in section 5.2.2.2, our interpretation was that there was no meaningful difference in conversational style between any of the voices. The differences between the voices in the perceptual evaluation were mainly related to their naturalness.

Hence, our conclusion is that utilising conversational speech data in HMM-based speech synthesis gives a better trade-off quality for synthesising propositional content wrapped in discourse markers and filled pauses than a conventional voice. But, we could not synthesise these conversational style utterances with as high quality as more conventional sentences. This suggests that in order to build HMM-based voices that can synthesise utterances typical for conversational speech it is better to build these voices from conversational data than to build them from a conventional data source. But, in order to reach the quality that can be achieved when synthesising conventional sentences (or preferably higher), additional modifications to the HMM-based framework are required.

The decision to evaluate the conversational style of the voices was partly motivated by the vast number of functions of different discourse markers and filled pauses. The conversational style was therefore selected to capture the general contribution of appropriately synthesised conversational speech phenomena on listeners' impression of the speech, rather than evaluating each function of discourse markers and filled pauses separately. The experiment with unit selection voices in section 5.6 exemplifies how specific functions of certain discourse markers and filled pauses can be evaluated. But, the most important motivation behind using conversational style as one of the evaluation criteria was to capture whether the speech modelling in HMM-based speech synthesis was capable of preserving the subtle phonetic detail that allow people to make a distinction between natural spontaneous or read aloud speech (see e.g. Blaauw (1994)). Evaluation of pragmatic function does not necessarily capture this (see the discussion in section 6.4). The result from the evaluation of the HMM-based voices in this section did not support that listeners identified a preserved distinction between the speaking styles in the source data. In contrast, in the evaluations of unit selection voices listeners readily identified which voice contained conversational speech, regardless of whether the voice was perceived as less natural (figure 4.9), equally natural (figure 5.8) or more natural (figure 5.8) than voices built without conversational speech.

5.3 Unit Selection Evaluations: Overview

In chapter 2 we argued that appropriately synthesised discourse markers and filled pauses wrapped around propositional content represents a key problem for synthesising conversational speech, because such utterances are frequently used in conversation to simultaneously express both propositional and non-propositional information. Appropriate synthesis would allow synthetic voices to express non-propositional information in conjunction with propositional content, as in the examples from the conversation with Johnny in chapter 3:

- agreeing about something with *oh yeah* in e.g. *oh yeah it's great exercise so*
- hesitating about something with filled pauses in e.g. *um, no I uh, uh I moved up for acting*
- being impressed by something with *wow* in e.g. *wow that's really cheap*
- or, asking someone for confirmation about something with *you know* in e.g. *whether successful or not I I aim for that, you know*.

Conventional synthetic voices are generally not designed to synthesise this type of non-propositional information in conjunction with the propositional content. Our hypothesis was that we could augment the conventional database of read aloud speech with speech from a recorded conversation to enable synthetic voices to express this combination of propositional and non-propositional information in a realistic manner. Hence, by blending conversational and read aloud speech data we aimed to build a synthetic voice that could synthesise a wide range of non-propositional information in conjunction with arbitrary propositional content.

In section 4.3 we described the developed unit selection blending method and the building of the different voices that we will evaluate. A summary of all the voices used in the evaluations are shown in table 5.2. We will go about evaluating these voices in three experiments that support complementary aspects of our approach to conversational speech synthesis:

1. Evaluating blending in section 5.4: When we utilise both conversational and read aloud speech, does the developed blending method contribute anything or can we build an equally good voice by treating the read aloud and conversational data as equivalent when building a voice?

2. Evaluating naturalness and conversational style in section 5.5: Does the inclusion of conversational speech data in a synthetic voice convey a general impression of a conversational style without a negative impact on naturalness?

3. Evaluating pragmatic function of conversational characteristics in section 5.6: Does the inclusion of conversational speech data in synthetic voices result in an improved ability to convey specific pragmatic functions?

The results from the evaluations will be discussed in connection with each experiment.

5.4 Evaluating Unit Selection Blending

The purpose of the experiment in this section was to show that coverage alone is not sufficient to produce good quality conversational speech synthesis. We cannot just include both conversational and read aloud data in a standard unit selection system and expect good quality speech. The techniques developed to segment conversational speech (in section 3.4.5) and control blending of conversational and read aloud speech (in section 4.3.3) are required in order to retain an acceptable level of naturalness. The synthetic voices evaluated in this section 5.4 experiment were described in section 4.3.4.1. The voice using a standard unit selection algorithm will be referred to as joh.16k.unit.baseline. The voice using the genre blending technique will be referred to as joh.16k.unit.blend. In summary, both voices were built with the same source data containing both the conversational and the read aloud speech data (see table 5.2). The differences between the voices were:

- In the joh.16k.unit.baseline voice both the read aloud and conversational speech were slowed down by 5%, since this was part of CereVoice's default audio processing. In the joh.16k.unit.blend, the read aloud data was slowed down by 5% and the conversational data was slowed down by 10%. This was done to address the difference in speaking rate between the read aloud and conversational data (see section 3.5.2.3).

- The joh.16k.unit.baseline forced alignment was made with the standard method described in section 2.3.1.1 where all the speech data was pooled in the training of the acoustic models. The joh.16k.unit.blend force aligned the read aloud and conversational data trained from genre-specific "adapted" acoustic models (see section 3.4.5).

- The joh.16k.unit.baseline voice used the standard target costs and weights in the CereVoice unit selection system. The joh.16k.unit.blend used a language model bias (see section 4.3.3) on top of the target cost to decide when units from the conversational data should be selected and when a backing-off to read aloud units should be done.

The experimental hypothesis was that the differences between the conversational and the read aloud data mean that uncontrolled use in a standard unit selection algorithm will lead to a significant loss of naturalness compared to a controlled blending approach.

5.4.1 Evaluation Design

The test material consisted of 10 news sentences selected from the material in Strom et al. (2007), and 10 conversational sentences selected from the material in section 4.2.6. Examples of the test material are shown in table 5.4. All the test sentences used in the evaluation are shown in appendix A, tables A.7 and A.8. The test sentences were synthesised with both voices and presented pairwise to the participants; randomised and in both orders, giving a total of 40 pairs. Twenty-three participants, both native and non-native English speakers, were paid to judge which utterance in a pair sounded more natural.

Voices	News text
joh.16k.unit.baseline	"Soldiers have lived a precarious existence within the posts, using state of the art listening devices and long range cameras to maintain round the clock surveillance."
joh.16k.unit.blend	"Soldiers have lived a precarious existence within the posts, using state of the art listening devices and long range cameras to maintain round the clock surveillance."
Voices	Conversation text
joh.16k.unit.baseline	"so let's see, but um, yeah, nothing exciting"
joh.16k.unit.blend	"so let's see, but um, yeah, nothing exciting"

Table 5.4: Examples of test sentence pairs for the evaluation in section 5.4. The text in the compared pairs are the same for both voices. Commas indicate where utterance internal silences were located. All the test sentences in the evaluation are shown in appendix A, tables A.7 and A.8.

5.4.2 Results

Figure 5.7 shows the data charts of the participants' judgements. The significance was tested with the binomial test. The results showed a significant ($p < 0.001$) loss of

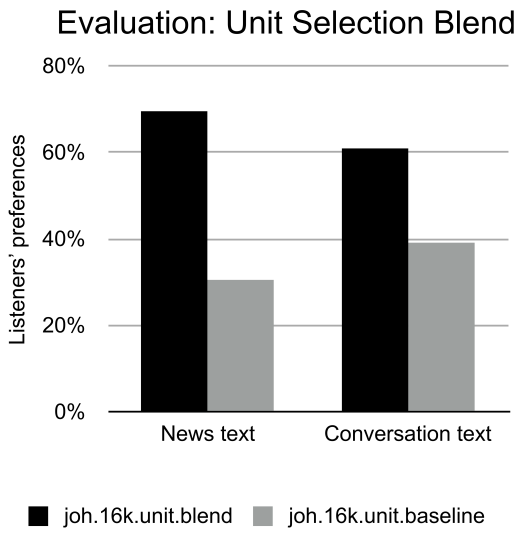

Figure 5.7: The bars show the percentages of the participants' preferences for naturalness when comparing the joh.16k.unit.baseline to joh.16k.unit.blend for text from news and conversation.

naturalness across both types of material for the joh.16k.unit.baseline voice compared to the controlled blending approach.

5.4.3 Conclusion

In chapters 2 and 3 we argued that blending conversational and read aloud data would allow us to alleviate the lack of phonetic coverage in spontaneous speech resources (see figure 3.4). The developed blending technique addressed the differences in language composition and phonetic properties found in the two data sources in chapter 3. The blending allowed us to utilise spontaneous speech phenomena directly from a recorded conversation, while utilising conventional speech resources of read aloud sentences to fill in the gaps in phonetic coverage.

The experiment in section 5.4 was conducted to confirm whether any blending was required. The result in section 5.4.2 confirmed that just including conversational speech together with read aloud speech in a standard unit selection system does not result in good quality synthetic speech. There seemed to be a contribution to this result from a degradation in forced alignment accuracy and more joins across read aloud and conversational speech, but there was no single factor identified that caused this loss of naturalness, and further analysis of the results in section 5.4 was not made. In sections 5.5 and 5.6 we will show that blending can be used to add conversational

characteristics to a conventional voice.

5.5 Evaluating Unit Selection Naturalness and Conversational Style

In section 2.3.2.2 we claimed that the challenge of blending in unit selection is to select conversational units to convey a conversational quality to the listeners, and to select read aloud units to maintain naturalness when there is a gap in the conversational coverage. The evaluation in section 5.4 confirmed that our developed techniques to blend conversational and read aloud data were required in order to not impair the quality of the synthetic speech. The experimental hypothesis tested in this section was whether the blending could be used to add conversational characteristics to a conventional voice and still maintain the same level of naturalness.

The voices used in the evaluation, joh.22k.unit.blend and joh.22k.unit.read, were described in section 4.3.3. The joh.22k.unit.read voice was built from the 103min read aloud data shown in table 5.1 with CereVoice's standard unit selection voice building method described in section 4.3.1. In the joh.22k.unit.blend voice the 75min conversational speech was added to the 103min read aloud data and the voice was built with the developed blending method described in section 4.3.4. The joh.22k.unit.blend voice included the speaking rate adjustment (see section 4.3.4), the style specific forced alignment (see section 3.4.5) and the language model control for selecting units from either conversational or read aloud speech (see section 4.3.3). An overview of the techniques and speech data used in the joh.22k.unit.blend and joh.22k.unit.read voices are shown in table 5.2.

Sections 5.5.1 and 5.5.2 contain a more detailed version of the collaborative work published in Andersson et al. (2010a). The evaluation design and analysis in Andersson et al. (2010a) were made by the current author. The design and analysis were discussed with the co-authors, in particular Kallirroi Georgila. The additional analysis in Section 5.5.2.1 was not part of the original publication, but was made by the author for this thesis.

5.5.1 Evaluation Design

The test sentences were randomly selected from a held-out set of the transcribed conversation. The discourse markers, filled pauses and disfluencies had been removed

from the transcript, and were replaced with predicted discourse markers and filled pauses. These predictions were generated by Kallirroi Georgila as described in section 4.3.5. To better evaluate the potential of predicting and synthesising a wide variety of types and placements of discourse markers and filled pauses (fillers), we restricted the selection of test sentences to contain the same filler sequence in at most two sentences. We selected 15 sentence pairs to synthesise:

- sentences with no fillers, referred to as NoFILL material, e.g.:
  "it's a different character for me"

- sentences with predicted fillers, referred to as FILL material, e.g.:
  "uh it's um [pause] a different character for me".

Both the FILL and NoFILL material were synthesised with the joh.22k.unit.read voice; henceforth referred to as FILL-joh.22k.unit.read and NoFILL-joh.22k.unit.read. In the listening test they were compared to FILL material synthesised with the joh.22k-.unit.blend voice; henceforth referred to as FILL-joh.22k.unit.blend. This gave us two test conditions: I) FILL-joh.22k.unit.blend vs. FILL-joh.22k.unit.read, and II) FILL-joh.22k.unit.blend vs. NoFILL-joh.22k.unit.read. Two examples of test sentence pairs are shown in table 5.5. All the sentence pairs used in the evaluation are shown in appendix A, table A.9. Table A.9 also shows which words were biased towards selecting units from the conversational speech data in the FILL-joh.22k.unit.blend utterances.

Thirty volunteering participants (both native and non-native English speakers) took part in the evaluation. The 15 sentence pairs for each of the two conditions were randomised and played to the participants in both orders. In total each participant listened to 60 sentence pairs of synthetic speech and were asked about their opinions on two different aspects:

- *Which utterance in the pair sounds more like in an everyday conversation (as opposed to e.g. someone reading from a script)?*

- *Which utterance in the pair sounds more natural (regardless if it sounds conversational or not)?*

The participants could express preference for either utterance in the pair ("A" or "B") or select a no-preference option ("Equal").

Material & Voice	Text	Pair No.
FILL-joh.22k.unit.blend	"uh it's um, a different character for me"	11 in
FILL-joh.22k.unit.read	"uh it's um, a different character for me"	fig. 5.9
FILL-joh.22k.unit.blend	"uh it's um, a different character for me"	11 in
NoFILL-joh.22k.unit.read	"it's a different character for me"	fig. 5.10

Table 5.5: Examples of test sentence pairs for the evaluation in section 5.5. Commas indicate where utterance internal silences were located. All the pairs in the evaluation are shown in appendix A, table A.9.

5.5.2 Results

Figure 5.8 shows the perceptual judgements for the two comparisons in the experiment: FILL-joh.22k.unit.blend compared to FILL-joh.22k.unit.read, and FILL-joh.22k.unit-.blend compared to NoFILL-joh.22k.unit.read.

The significance of the result was tested with the binomial test. The times when participants expressed no preference were removed before calculating the results. The FILL-joh.22k.unit.blend utterances were perceived as significantly ($p < 0.001$) more conversational than the FILL-joh.22k.unit.read utterances. The FILL-joh.22k.unit-.blend utterances were also perceived as significantly ($p < 0.001$) more natural than the FILL-joh.22k.unit.read utterances. This means that it is not sufficient to just insert discourse markers and filled pauses in text, but it is essential to have appropriate realisations of discourse markers and filled pauses in the voice, otherwise naturalness is negatively affected.

The FILL-joh.22k.unit.blend utterances were perceived as significantly ($p < 0.001$) more conversational than the NoFILL-joh.22k.unit.read utterances. The FILL-joh.22k-.unit.blend utterances and the NoFILL-joh.22k.unit.read utterances were not perceived as significantly ($p = 0.28$) different in terms of how natural they sounded. This means that we can include conversational speech in synthesis to achieve a more conversational style without decreasing the general naturalness.

5.5.2.1 Naturalness and Conversational Style

In figure 5.8 the perceptual judgements were collapsed over all utterances. These collapsed judgements also corresponded well to the judgements for individual utterances, except for the comparison of naturalness for the FILL-joh.22k.unit.blend and NoFILL-joh.22k.unit.read utterances. In figure 5.8 this comparison showed that, on average,

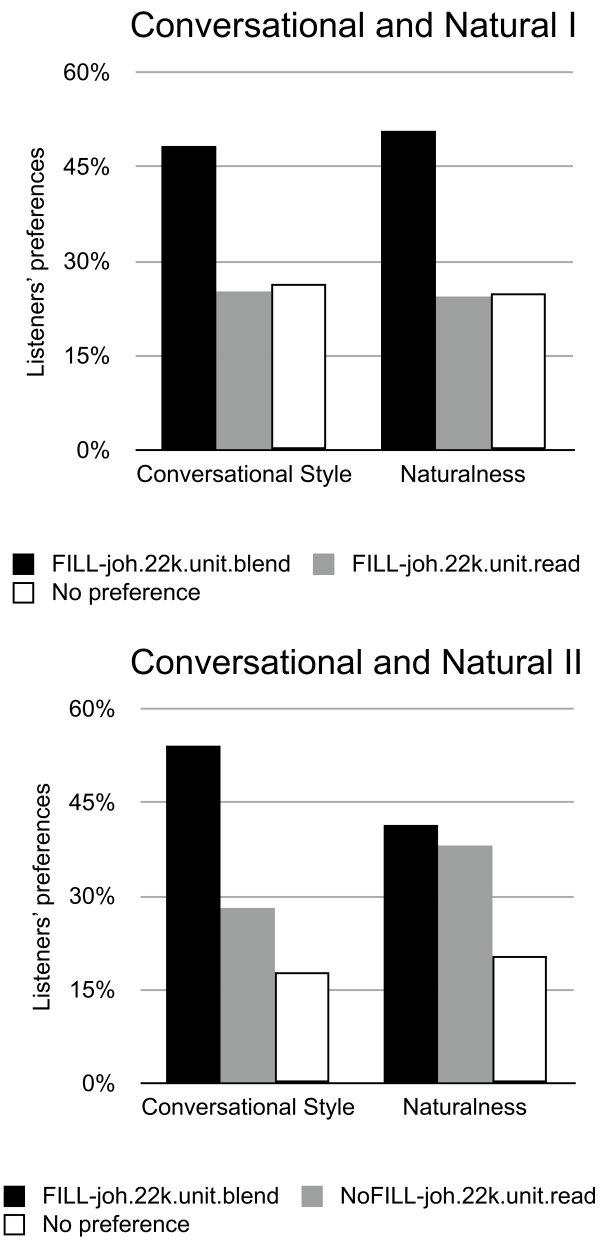

Figure 5.8: Percentage of perceptual judgements for “Conversational” and “Natural” quality of synthetic speech when comparing the FILL-joh.22k.unit.blend to the FILL-joh.22k.unit.read and NoFIll-joh.22k.unit.read. The “No preference” shows the percentage of listeners who expressed no preference for either voice.

they sounded equally natural, but the results for individual utterances in figure 5.9 show that the FILL-joh.22k.unit.blend utterances sometimes sounded more natural, and sometimes less natural, than the NoFILL-joh.22k.unit.read utterances. In this section we will analyse why some utterances were perceived by the participants as more or less natural.

The blending of read aloud and conversational speech was designed to alleviate the lack of phonetic coverage in the conversational speech data. During synthesis of the FILL-joh.22k.unit.blend utterances the selection of units was biased towards selecting units from the conversational speech, if words in the input text existed in the conversational speech data (see section 4.3.3). Only two out of the fifteen evaluation utterances from the FILL-joh.22k.unit.blend consisted entirely of units from the conversational speech. One of these utterances was perceived by the participants as being much more natural than the corresponding NoFILL-joh.22k.unit.read utterance (utterance number 05 in fig. 5.9), and the other was perceived as less natural (utterance number 03 in fig. 5.9). Neither the NoFILL-joh.22k.unit.read nor the FILL-joh.22k.unit.blend utterances (03 and 05) had any prominent concatenation errors, and the difference between the perceived naturalness is likely to have been due to utterance level prosody. The rest of the FILL-joh.22k.unit.blend utterances all contained units from both read aloud and conversational speech. Table 5.6 and table 5.7 show blending of conversational and read aloud speech in utterances from the FILL-joh.22k.unit.blend that were perceived by the participants as more (table 5.7) or less (table 5.6) natural than the utterances from the NoFILL-joh.22k.unit.read.

When comparing the utterances in table 5.6 and table 5.7 there was more blending in the utterances that were perceived as less natural, in particular there were longer (more than one) word sequences of read aloud units in the utterances that were perceived as less natural. However, utterance 11 in table 5.6 did not contain many read aloud units, the problem in this utterance was that the *-er* in *character* resulted in a concatenation error, whereas the corresponding NoFILL-joh.22k.unit.read utterance did not have any concatenation errors. Similarly, in utterance 14 in table 5.7 the FILL-joh.22k.unit.blend utterance contained quite a few units from read aloud speech, but the corresponding NoFILL-joh.22k.unit.read utterance had a prominent concatenation error in *apparently*.

The conclusions drawn from this qualitative analysis of naturalness is that some blending did not have a negative impact on speech quality, but too much blending made the speech quality less coherent, and sounded like speech units spliced together

Utt. No.	Material & Voice	Text
00	FILL-joh.22k.unit.blend	*you know uh, I wasn't* **too embarrassed** *to say* **that's di***sgu***sting**
	NoFILL-joh.22k.unit.read	**I wasn't too embarrassed to say that's disgusting**
06	FILL-joh.22k.unit.blend	*uh then so, I* **just wanna throw something**
	NoFILL-joh.22k.unit.read	**then I just wanna throw something**
09	FILL-joh.22k.unit.blend	*uh you know, but as far as* **getting** *out the* **theatres it has not done well**
	NoFILL-joh.22k.unit.read	**but as far as getting out the theatres it has not done well**
11	FILL-joh.22k.unit.blend	*uh it's um,* **a** *different* **charact***er for me*
	NoFILL-joh.22k.unit.read	**it's a different character for me**

Table 5.6: Examples of utterances shown in figure 5.9 where the FILL-joh.22k.unit-.blend utterances sounded **less** natural than the NoFILL-joh.22k.unit.read utterances. For the FILL-joh.22k.unit.blend utterances the bold faced text shows where units were selected from read aloud speech, and the italic text shows where units were selected from conversational speech. For the NoFILL-joh.22k.unit.read utterances all units came from read aloud speech, and the text is therefore shown in bold.

from different utterances. Blending was also selected as more natural than having concatenation artefacts by listener. Additionally, whereas the FILL-joh.22k.unit.blend utterances were sometimes more natural and sometimes less natural than the NoFILL-joh.22k.unit.read, figure 5.10 shows that the FILL-joh.22k.unit.blend utterances were consistently perceived as having a more conversational style. Hence, blending could be used to synthesise high quality speech with a distinctly conversational style.

5.5.3 Conclusion

The experiment in section 5.5 compared a synthetic voice (joh.22k.unit.read) built from conventional read aloud sentences to a synthetic voice (joh.22k.unit.blend) built from a combination of both the read aloud sentences and speech from a spontaneous conversation. Table 5.2 summarises the synthesis methods and speech data used for the two voices.

The joh.22k.unit.read and joh.22k.unit.blend voices were compared to test the hypothesis of this thesis that we can utilise conversational speech to add conversational characteristics to a conventional "neutral" synthetic voice and still maintain the same level of naturalness as the conventional voice. The results from the perceptual evaluation in section 5.5.2 supported that hypothesis. The evaluation compared the joh.22k-.unit.blend and joh.22k.unit.read voices under two conditions. The two conditions are exemplified in table 5.5. In the first condition we tested the joh.22k.unit.blend voice's

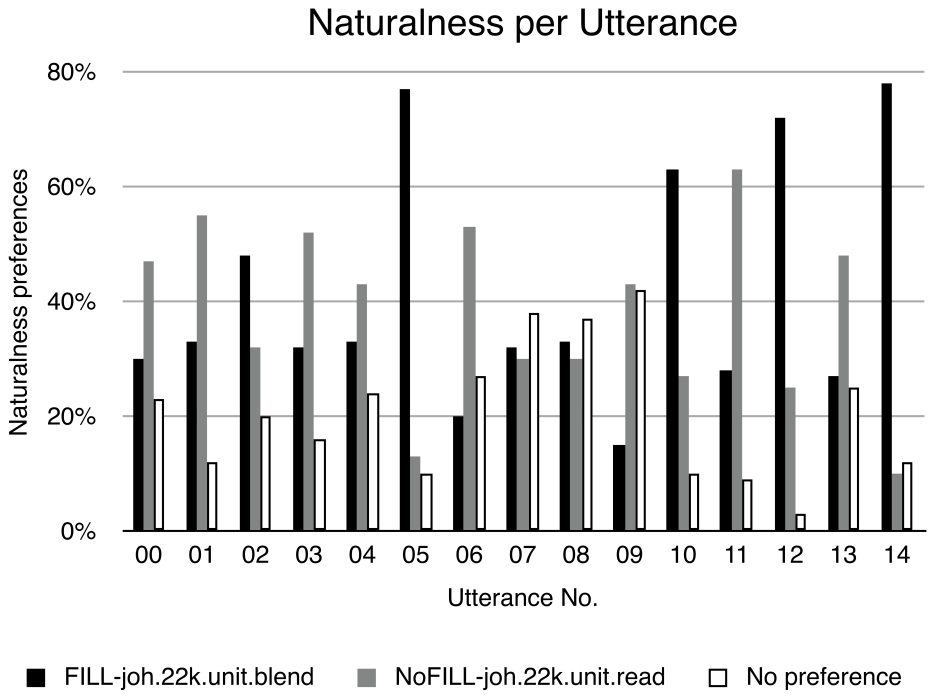

Figure 5.9: Perceptual judgements of "Naturalness" for individual utterances, when comparing the FILL-joh.22k.unit.blend utterances to the NoFILL-joh.22k.unit.read utterances. The "No preference" shows percentage of listeners who expressed no preference for either voice in an utterance pair.

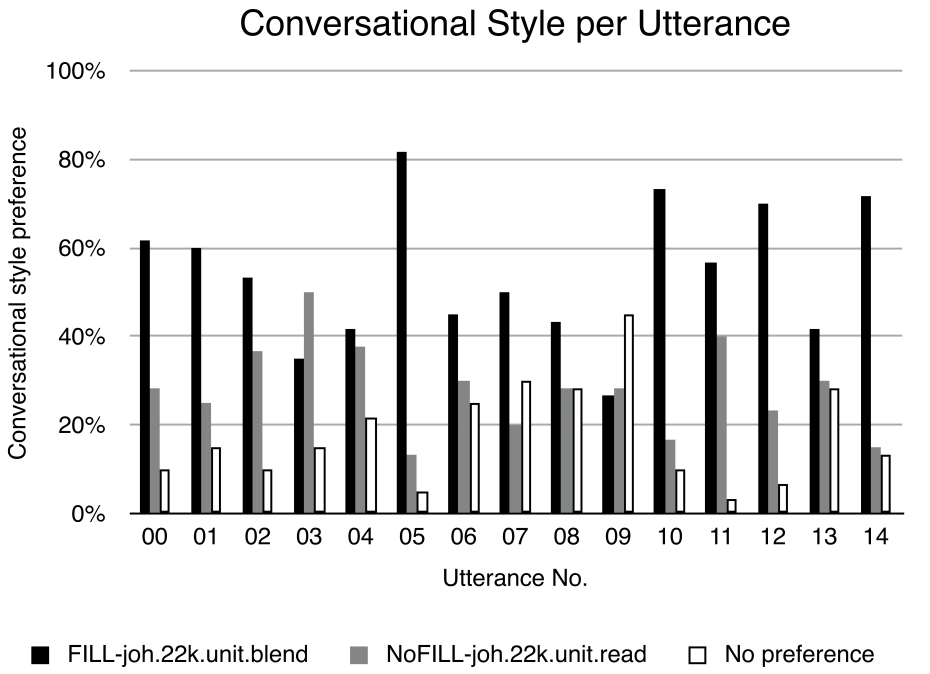

Figure 5.10: Perceptual judgements of "Conversational style" for individual utterances, when comparing the FILL-joh.22k.unit.blend utterances to the NoFILL-joh.22k.unit.read utterances. The "No preference" shows percentage of listeners who expressed no preference for either voice in an utterance pair.

Utt. No.	Material & Voice	Text
10	FILL-joh.22k.unit.blend	*uh you know but um, you know, I went there, and so uh, I was there for a few* **weeks**
	NoFILL-joh.22k.unit.read	**but, I went there, and I was there for a few weeks**
12	FILL-joh.22k.unit.blend	**long** *story short it's garbage, my god, um, it is* **garb***age*
	NoFILL-joh.22k.unit.read	**long story short it's garbage, my god it is garbage**
14	FILL-joh.22k.unit.blend	**apparently,** *yeah, I know way too much* **about,** *like, the* **sex stuff** *here in* **America**
	NoFILL-joh.22k.unit.read	**apparently I know way too much about the sex stuff here in America**

Table 5.7: Examples of utterances shown in figure 5.9 where the FILL-joh.22k.unit-.blend utterances sounded **more** natural than the NoFILL-joh.22k.unit.read utterances. For the FILL-joh.22k.unit.blend utterances the bold faced text shows where units were selected from read aloud speech, and the italic text shows where units were selected from conversational speech. For the NoFILL-joh.22k.unit.read voice all units came from read aloud speech, and the text is therefore shown in bold.

ability to synthesise frequent conversational speech phenomena; discourse markers and filled pauses wrapped around propositional content, compared to a conventional voice (joh.22k.unit.read). In the second condition, we tested to what extent the joh.22k.unit-.blend voice could synthesise material with discourse markers and filled pauses as well as the joh.22k.unit.read voice could synthesise material without these frequent conversational speech phenomena. In both conditions, the joh.22k.unit.blend voice was perceived as more conversational than the joh.22k.unit.read voice. In none of the conditions was the joh.22k.unit.blend voice perceived as less natural than the conventional joh.22k.unit.read voice. Hence, the results showed that the perceived conversational character of the joh.22k.unit.blend voice was due to the voice's ability to synthesise conversational characteristics in a more natural manner than the conventional joh.22k-.unit.read voice.

The evaluation was unconventionally designed in that the joh.22k.unit.blend voice contained substantially more data than the joh.22k.unit.read voice. But, this was the point: to test if we could add behaviour by augmenting the conventional voice with speech from a spontaneous conversation. It is possible that comparing the joh.22k-.unit.blend voice to a voice with acted sentences with the same linguistic content, and hence very similar size and phonetic coverage, would give a different result: an acted voice could sound better because it was more similar to neutrally read sentences (bad acting), it could sound worse because it is difficult to act or read aloud transcribed conversations (bad acting), or it could sound the same because our actor managed to make the utterances sound like in a spontaneous conversation (good acting). Hence,

any result of naturalness or conversational style from such a comparison would only depend on the quality of the actor. Our aim was to enrich the limited expressiveness of conventional voices by adding a controlled set of speech from a spontaneous conversation. The evaluation was designed to test this approach. Contrasting spontaneous speech with acted speech was considered outside the scope of the thesis.

In section 5.6 we will investigate whether the improved ability to synthesise discourse markers and filled pauses wrapped around propositional content also results in an improved capacity of conveying pragmatic information to listeners.

5.6 Evaluating Function of Conversational Characteristics

The experiment in section 5.5 showed that the joh.22k.unit.blend voice added a general conversational quality to the synthetic speech compared to the joh.22k.unit.read voice. The experiment in this section was conducted to investigate whether the general conversational quality also resulted in an improved ability to convey specific pragmatic functions.

The voices used in the evaluation, joh.22k.unit.blend and joh.22k.unit.read, were the same as in section 5.5. The joh.22k.unit.read voice was built from the 103min conversational data shown in table 5.1 and the joh.22k.unit.blend voice was built from both the 103min read aloud data and the 75min conversational data. Table 5.2 contains a summary of the methods and speech data used for the two voices. A more detailed description of the voices can be found in section 4.3.4.

In section 2.2 we described the discourse markers and filled pauses that we have focused this thesis on. The discourse markers and filled pauses consist of words and expressions that are frequently used in conversation to express a wide range of non-propositional information, for example:

- collocations with *yeah*, e.g. *oh yeah*, are often used to express agreement (Jurafsky et al., 1998)

- the filled pauses are often used to express hesitation (Clark and Fox Tree, 2002).

As described in section 2.2.2.3, previous research has shown that the filled pauses affect listeners' perception of the speaker's certainty about a topic, in both natural (Brennan and Williams, 1995) and synthetic speech (Lasarcyk and Wollerman, 2010).

Therefore we designed an experiment to investigate whether the better synthesis of propositional content wrapped in filled pauses and discourse markers in the joh.22k-.unit.blend voice could be utilised to make a pragmatic contrast in certainty. Specifically, we investigated whether the agreement of *yeah* and the hesitation of filled pauses could be used to synthesise a pragmatic contrast and convey certainty or uncertainty about a topic.

The experimental hypothesis was that a blended conversational synthetic voice conveys pragmatic elements of conversational speech more effectively than a conventional synthetic voice. The experiment is limited to a specific pragmatic function; the conveyance of (un)certainty. As such it can only accept or reject the hypothesis in this domain.

5.6.1 Evaluation Design

For this evaluation we designed sentences that would potentially convey different pragmatic functions. Seven utterances with initial *yeah*-sequences, e.g. *right yeah, about two years ago* were designed to convey certainty (CERT). Seven utterances with initial *um*-sequences, e.g. *well, you know um, about two years ago*, were designed to convey uncertainty (UNCERT). All sentences used in the evaluation are shown in appendix A, table A.10.

The CERT and UNCERT material were synthesised with the joh.22k.unit.blend and joh.22k.unit.read voices. Two natural speech samples for each of CERT and UNCERT were included for reference. The natural samples were selected to resemble the design of the CERT and UNCERT material and express certainty or uncertainty:

- *yeah, in*[1] *Monday they're buying*
- *yeah, no I can talk without needing a break*
- *um, no I uh, uh I moved up for, acting*
- *I just saw um, uh um, a version of a film that I did, um, in Thai*

Twenty-three participants, both native and non-native English speakers, were paid to take part in the experiment. The participants were requested to judge the certainty of each speech sample on a MOS scale from *1 very certain* to *5 very uncertain*. The

[1] When transcribing, it can be tempting to go with the grammatical choice, in this case *on*, but in this case the author's interpretation was that Johnny said wrong.

general difference in naturalness for conversational material between the joh.22k.unit-.blend and joh.22k.unit.read voice was established in section 5.5.2 and confirmed for the CERT and UNCERT material through conventional MOS of naturalness (from *1 completely unnatural* to *5 completely natural*).

5.6.2 Results

Boxplots for the perceived certainty and uncertainty are shown in figures 5.11 and 5.12. The significance for the perceived certainty and uncertainty was tested with pairwise Mann-Whitney, and Bonferroni correction of significance levels. Table 5.8 shows the results of these tests.

The difference between perceived certainty in utterances with CERT or UNCERT material is significant for all voices. The differences between joh.22k.unit.blend and joh.22k.unit.read voices are significant for CERT, but not UNCERT material. Hence, the joh.22k.unit.blend voice achieved the pragmatic objective of communicating changes in certainty, although not as effectively as natural speech. The joh.22k.unit.read voice was less effective in communicating the pragmatic function.

The perceived naturalness of the voices was also tested. The median value for the natural speech samples (both CERT and UNCERT material) was 5 (i.e. the natural speech was perceived as natural). The median value for the joh.22k.unit.blend CERT and UNCERT material was 4, and the median value for joh.22k.unit.read CERT and UNCERT material was 2. The significance of these differences was tested with pairwise Mann-Whitney, and Bonferroni correction of significance levels. The natural material was significantly more natural than both joh.22k.unit.blend ($p < 0.001$) and joh.22k.unit.read ($p < 0.001$) material. The joh.22k.unit.blend material was significantly more natural than joh.22k.unit.read material ($p < 0.001$).

5.6.3 Conclusion

The experiment in section 5.5 found that the blended voice (joh.22k.unit.blend) added a general conversational quality to the synthetic speech. In the experiment described in this section we investigated whether this conversational quality resulted in an improved capability of conveying specific pragmatic functions.

The different discourse markers and filled pauses are associated with many different non-propositional functions (see section 2.2). In line with previous research (Brennan and Williams, 1995; Lasarcyk and Wollerman, 2010), we therefore chose to

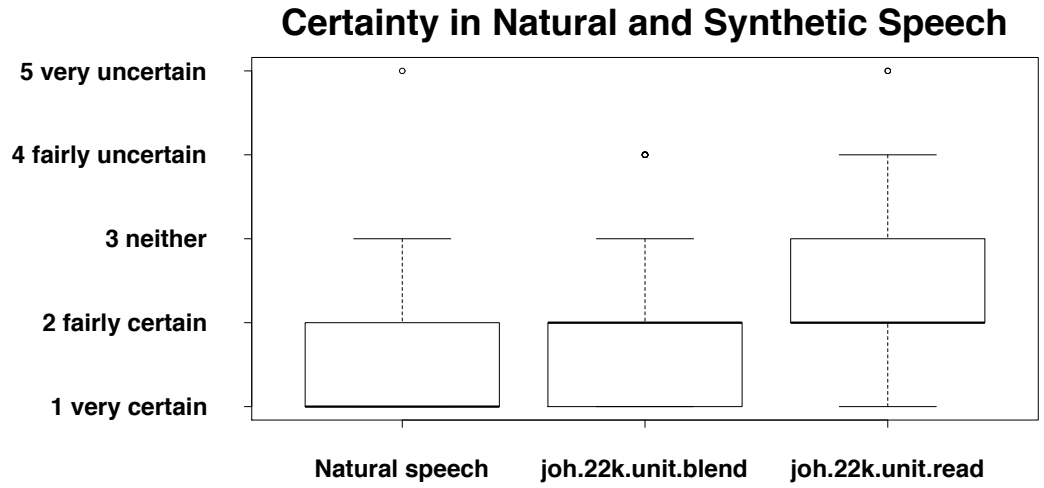

Figure 5.11: The figure shows participants' judgements of certainty in utterances with CERT material, e.g. *"yeah, a vast majority of the members"*

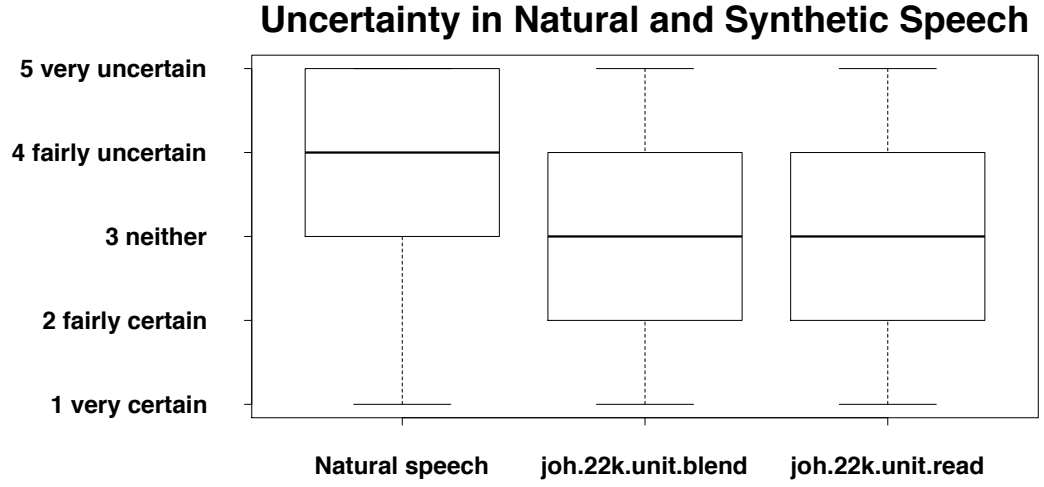

Figure 5.12: The figure shows participants' judgements of certainty in utterances with UNCERT material, e.g. *"you know um, uh, a vast majority of the members"*

evaluate pragmatic function in one specific dimension: (un)certainty. We then designed and synthesised sentences that would potentially convey certainty (CERT), e.g. *right yeah, about two years ago*, or uncertainty (UNCERT), e.g. *well, you know um, about two years ago*. Natural speech samples that were hypothesised to convey (un)certainty in a similar manner were also included in the evaluation.

The results showed that all voices conveyed a difference in certainty between the CERT and UNCERT material (see figures 5.11 and 5.12). The results also showed that the blended conversational voice could synthesise a more prominent contrast in (un)certainty than the conventional voice, although not as prominent as in the natural speech. The conclusion drawn from these results is that the blended conversational voice was more effective in conveying pragmatic contrast than the conventional voice.

5.7 Conclusion

The aim of this thesis is to produce synthetic speech that can express conversational characteristics in a natural manner. In this chapter we conducted perceptual experiments to determine whether the synthetic voices built in chapter 4 could produce such speech.

In chapter 2 and chapter 3 we argued that the key challenge for creating conversational synthetic speech lies in improving synthesis of frequent conversational speech phenomena. Specifically, we argued that we should improve synthesis of discourse markers and filled pauses integrated with propositional content, because that would allow synthetic voices to express a wide range of non-propositional information in

Experimental Hypothesis (H1)	p-value
CERT is more certain than UNCERT when both types of material are synthesised with joh.22k.unit.blend	<0.001
CERT is more certain than UNCERT when both types of material are synthesised with joh.22k.unit.read	<0.001
CERT is more certain than UNCERT when both types of material are natural speech	<0.001
CERT synthesised with joh.22k.unit.blend is more certain than CERT synthesised with joh.22k.unit.read	<0.001
CERT from natural speech is more certain than CERT synthesised with joh.22k.unit.blend	<0.001
CERT from natural speech is more certain than CERT synthesised with joh.22k.unit.read	<0.001
UNCERT synthesised with joh.22k.unit.blend is less certain than UNCERT synthesised with joh.22k.unit.read	0.86
UNCERT from natural speech is less certain than UNCERT synthesised with joh.22k.unit.blend	<0.001
UNCERT from natural speech is less certain than UNCERT synthesised with joh.22k.unit.read	<0.001

Table 5.8: The Bonferroni corrected p-values from the statistical analysis. The significance of each hypothesis was tested with Mann-Whitney.

conjunction with the propositional content, e.g. certainty or uncertainty about a topic, in a manner similar to how people express it in spontaneous conversations. The perceptual evaluations conducted in this chapter therefore tested our conversational voices' ability to synthesise discourse markers and filled pauses wrapped around propositional content. The example from the conversation in chapter 3 illustrates this frequent utterance structure, where the propositional content is bold faced and discourse markers and filled pauses are in italics: "*yeah exactly and even like uh* **I'll go and see bad movies that I know will be bad** *um just* **to see why they're so bad**".

The different evaluations targeted different aspects of the synthetic voices. In total, we evaluated three aspects of the synthetic speech: the general quality (naturalness), the conversational quality (conversational style), and the ability to convey specific pragmatic functions (certainty or uncertainty) to the listeners. The general hypothesis tested was that we can utilise conversational speech to add conversational characteristics to synthetic voices without a negative impact on the naturalness. Table 5.2 contains a summary of the methods and speech data used for the different voices in the experiments.

The experiment in section 5.2 evaluated the naturalness and conversational style of the HMM-based synthetic voices. The evaluation was designed to test a) which voice sounded more natural, and b) whether we could preserve two distinct speaking styles in the HMM-based voices: read aloud or conversational. The naturalness and conversational style were evaluated by two separate groups of listeners. Three HMM-based voices were tested:

- joh.16k.hts.spon: built from 75min speech from a spontaneous conversation.
- joh.16k.hts.read: built from 103min neutrally read aloud sentences.
- joh.16k.hts.blend[spon|read]: built from both the 75min conversational and the 103min read aloud data. A blending technique was applied to allow training and synthesis of two speaking styles (see section 4.2.5). When synthesising a "spontaneous" or "read aloud" speaking style with this voice, we refer to it as joh.16k.hts.blendspon or joh.16k.hts.blendread.

The results in figure 5.1 showed that the joh.16k.hts.spon was more natural than the joh.16k.hts.read, when synthesising held-out conversational material. The phonetic analysis in section 4.2.6 supported our impression that the most prominent differences in quality were in the more natural synthesis of discourse markers and filled

pauses. But, the results from the blended voice in figure 5.2, joh.16k.hts.blendspon and joh.16k.hts.blendread, showed that we could not synthesise held-out conversational material with as high quality as more conventional material.

The results in figure 5.1 also showed that the joh.16k.hts.spon was perceived as having a more conversational style than the joh.16k.hts.read, whereas the results for the blended voice in figure 5.2 did not support our hypothesis that we could preserve two speaking styles with the applied blending technique. There was a strong correlation found between listeners judgement of naturalness and conversational style (see section 5.2.2.2), and taken together with the qualitative analysis in 5.2.2.1, our conclusion was that there was no meaningful difference in speaking style between any of the HMM-based voices. The differences between the voices were mainly related to their naturalness. This means that the HMM-based voices were not natural enough for listeners to identify the subtle phonetic detail that allows listeners to make a distinction between natural read aloud and spontaneous speech in Blaauw (1992, 1994).

The evaluation design for the HMM-based voices did not allow us to draw conclusion as to whether the blending resulted in a better quality than either of the style-dependent voices. Hence, it did not directly test the hypothesis of this thesis: that blending can be used to add conversational characteristics to a conventional synthetic voice, while maintaining the same level of naturalness as the conventional voice. However, the results from the style-dependent joh.16k.hts.spon and joh.16k.hts.read voices showed that despite less material and less phonetic coverage the joh.16k.hts.spon sounded better than the joh.16k.hts.read. This demonstrated the generalisation power of HMM-based speech synthesis, where domain coverage made a positive contribution, while the lack of general coverage could be compensated for. This is very similar to what the blending was intended to achieve. Therefore, we did not consider it worthwhile to re-design the experiment to test the blending. There are other areas than blending that need to be improved in HMM-based speech synthesis in order to compete with the quality of human voices, where the most obvious area is the vocoder's limitations for synthesising different voice qualities (Cabral et al., 2008; Silén et al., 2009).

Unit selection has the ability to preserve local phonetic detail of natural speech, but also has an associated weakness in compensating for gaps in the phonetic coverage. Blending is therefore a potential solution to preserving the local phonetic detail of conversational speech phenomena while compensating for the gaps in phonetic coverage with read aloud sentences. The evaluation of the unit selection voices therefore addressed the hypothesis of this thesis more directly than the HMM-based experiments.

The evaluation was made in a series of three experiments:

1. In section 5.4, we evaluated whether the developed blending technique was required. This was evaluated by comparing two voices built from both the 75min conversational data and the 103min read aloud data shown in table 5.1. The difference between the voices was that one was built with the developed blending method and the other was built with the standard unit selection method. The results showed that the voice using the standard unit selection method was perceived as less natural than the voice using the blending method, for both news sentences and held-out conversational material. The conclusion drawn was therefore that the developed blending method was required in order to not impair the quality of the synthetic speech.

2. In section 5.5, we evaluated whether the blended voice was perceived as a) having a more conversational style than the conventional voice, and b) not being less natural than the conventional voice. The blended voice (joh.22k.unit.blend) contained both the 75min conversational data and the 103min read aloud data. The conventional voice (joh.22k.unit.read) was built from the 103min read aloud data. The results in figure 5.8 showed that the joh.22k.unit.blend voice was perceived as more conversational and not less natural than the joh.22k.unit.read voice. Hence, the results supported the hypothesis of this thesis: we can use speech from a spontaneous conversation to add conversational characteristics to conventional voices without impairing the naturalness.

3. In section 5.6, we evaluated whether the blended joh.22k.unit.blend voice could convey specific pragmatic information, certainty or uncertainty about a topic, in a more convincing manner than the conventional joh.22k.unit.read voice. The test material was designed so that the test sentences only differed in the utterance initial discourse markers or filled pauses. An example to express certainty is: *right yeah, about two years ago*, and an example to express uncertainty is: *well, you know um, about two years ago*. This was testing whether the more natural synthesis of these speech phenomena also meant that the pragmatic information could be communicated more efficiently. The results in figures 5.11 and 5.12 showed that both the joh.22k.unit.blend and joh.22k.unit.read voice conveyed a pragmatic contrast in the synthesised material, but that the joh.22k.unit.blend voice communicated a greater contrast. The conclusion drawn is that by improving synthesis of discourse markers and filled pauses wrapped around proposi-

tional content we can make synthetic voices better at communicating pragmatic contrasts.

The evaluations in this chapter of the synthetic voices demonstrated the effect on listeners of using speech from a spontaneous conversation in unit selection and HMM-based speech synthesis systems. The evaluation of the unit selection voices directly supported the hypothesis of this thesis by showing that we could make a conventional voice exhibit more conversational characteristics by augmenting it with carefully selected speech from a spontaneous conversation. The experiments with the HMM-based voices did not directly test the thesis hypothesis, but the results supported the more general hypothesis that speech from a spontaneous conversation in the voices improves synthesis of conversational speech phenomena. In chapter 6 we will make a concluding discussion of our approach to conversational speech synthesis.

Chapter 6

Conclusion

The aim of this thesis was to create synthetic voices that could convey an impression of natural-sounding conversational characteristics. In the previous chapters we have described how we approached that problem by blending speech from a spontaneous conversation with conventional read aloud speech resources. In this chapter we will make a concluding discussion of our approach, methods, and achieved results.

6.1 Conversational Speech Synthesis

Conventional synthetic voices generally focus on synthesising intelligible and natural-sounding propositional information. These qualities make the voices well suited to read aloud driving directions in a GPS system. But, there are other applications for synthetic voices that have other quality requirements on the voices. The believable characters in, for example, Traum et al. (2008) require voices that allow the characters to partake in conversations in a realistic manner. Currently, the realism of these conversations is limited partly by the characters' inability to express themselves in a natural conversational manner outside a set of pre-recorded prompts. In order to extend these characters' ability to express themselves, their voices need to be able to synthesise a richer variety of the speech phenomena found in human conversations.

In section 2.2 we described how utterances in a conversation are often constructed by wrapping propositional content in discourse markers and filled pauses. This utterance structure is frequently used in spontaneous conversation to express a wide range of non-propositional information together with the propositional content, as exemplified in the samples from the conversation with Johnny in chapter 3:

- Expressing agreement or certainty about a topic with *yeah* or *oh yeah* as in: *oh yeah it's great exercise so*.

- Expressing hesitation or uncertainty with the filled pauses, e.g. *um, no I uh, uh I moved up for acting*.

- Asking for confirmation from the conversation partner by ending the utterance with *you know* in e.g. *whether successful or not I I aim for that, you know*.

Conventional synthetic voices are generally not designed to synthesise this integration of propositional and non-propositional information. Synthesising these types of utterances in a natural conversational manner therefore represents a key problem for enabling synthetic voices to express themselves more naturally in conversation.

6.1.1 The Blending Approach

Unit selection and HMM-based speech synthesis frameworks build voices by modelling the phonetic properties from recordings of natural speech. Therefore, we aimed to capture the natural phonetic properties of conversational speech phenomena by utilising speech from an actual spontaneous conversation to build unit selection and HMM-based synthetic voices. However, the unit selection and HMM-based speech synthesis frameworks require phonetic coverage in order to synthesise high quality speech. The analysis of the conversation in section 3.5 showed that it contained a substantial amount of discourse markers and filled pauses wrapped around propositional content, which is required in order to synthesise them with unit selection or HMM-based speech synthesis methods. But, figure 3.4 showed that it is not feasible to achieve phonetic coverage by recording spontaneous conversation alone. Therefore, we developed methods to blend speech from a conversation with conventional read aloud sentences. These read aloud sentences were pre-selected for providing phonetic coverage. This approach would allow the synthetic voices to preserve the phonetic properties of frequent conversational speech phenomena, while maintaining consistently high quality synthesis by boosting the phonetic coverage with read aloud sentences. The resulting voices would then be able to synthesise a rich variety of non-propositional information in conjunction with arbitrary propositional content.

To achieve our goal of conversational speech synthesis through the blending approach we addressed the four research questions stated in section 1.4:

- How to obtain spontaneous conversations under the controlled conditions required for building high quality synthetic voices.
- How to constrain the rich variety of speech phenomena in a spontaneous conversation to create a controlled dataset of conversational utterances from which we can automatically build high quality synthetic voices in conventional speech synthesis systems.
- To what extent can we alleviate the lack of control over phonetic coverage in spontaneous speech resources by blending conventional pre-selected and neutrally read aloud data with data from a conversation.
- To what extent does the inclusion of conversational speech in synthetic voices influence listeners' impression of conversational speaking style and pragmatic meaning of synthetic utterances.

The hypothesis was that by augmenting the conventional database of neutrally read aloud speech with conversational data, we could add conversational characteristics to conventional synthetic voices without causing a negative effect on quality.

6.2 Challenges

In this section we will summarise the challenges that we faced when addressing our four research questions.

6.2.1 Obtaining Conversational Speech

The read aloud sentences that are conventionally used for speech synthesis are recorded in a recording studio. This gives speech recorded in a controlled environment under consistent conditions without background noise. To obtain high quality recordings of conversational speech, our recorded conversations in chapter 3 took place in the same recording studio environments as the recordings of the read aloud sentences. The matching recording conditions of the conversational and read aloud speech were also intended to facilitate the blending of the two data sources.

Our recording conditions met the most fundamental conditions for spontaneous conversation, as outlined in section 2.1: face-to-face, real-time and non-scripted (Clark, 1996). Any negative effect from the artificial constraints of recording spontaneous

conversation in a studio with a paid voice talent was therefore hypothesised to be negligible for the purpose of obtaining conversational speech for synthesis. The question was rather a matter of how the interpersonal relation between two strangers would affect the conversation: can we record enough conversation for use in speech synthesis, or will the conversation become forced and awkward after a few minutes?

A voice talent is cast for their particular speech characteristics. When recording conversation, the dynamics between the participants need to be considered. Just as the recording of Roger, a middle aged man, gives a different synthetic voice than the recordings of Heather, a young woman, the conversation between two middle aged men is likely to be different from the conversation between two young women. In our work we cast the author as conversation partner to the voice talents in chapter 3. As a result we managed to elicit about seven hours of spontaneous conversation with Johnny, which we turned into almost one and a half hours of conversational data to use for building synthetic voices (see chapter 3).

Hence, recording spontaneous conversation in a studio is a straightforward method for obtaining conversational speech for synthesis. The nature of the conversation will depend on the dynamics between the cast conversation participants, rather than just an individual voice talent.

6.2.2 Controlling Spontaneity for Speech Synthesis

Unit selection and HMM-based speech synthesis systems build synthetic voices from recordings of speech, where each audio file has a corresponding orthographic transcript that precisely matches the audio. This precise match is required in order to build high quality synthetic voices. Therefore, to get this precise match also for speech from a spontaneous conversation, the conversation was transcribed manually. A spontaneous conversation contains an abundance of speech phenomena that are currently not modelled well in speech synthesis, e.g. heavily reduced pronunciations, mispronunciations, word fragments, interrupted utterances, mumbling and laughter. Utterances that contained these speech phenomena were excluded. As a result we obtained a controlled set of conversational speech data that contained mainly the speech phenomena that we focused on: discourse markers, filled pauses and propositional content.

Even with a data set of conversational speech where the included speech phenomena were controlled, there remained challenges. The speech needed to be segmented into phoneme-sized segments and the discourse markers and filled pauses needed to

be represented in a manner that would allow them to be synthesised with our synthetic voices. We showed in section 3.4.5 that the segmentation of the conversational data could be made sufficiently accurate for speech synthesis by modifying the forced alignment method in our speech synthesis system. The alternative of manually segmenting the conversational speech data was rejected on the basis that it is too resource intensive.

The most important criterion for representing speech phenomena for synthesis is to what extent the representation enables synthesising the speech phenomena, i.e. allows generating the phonetic properties of the speech phenomena for unrecorded material. The majority of discourse markers and filled pauses consists of a limited set of expressions that occur in the vicinity of a phrase boundary, shown for our data in table 3.4. In section 2.2 we therefore argued that the shallow linguistic features phoneme sequence and utterance position, that were automatically extracted in our speech synthesis systems, would be sufficient to preserve the phonetic properties and associated prototypical pragmatic function also for the discourse markers and filled pauses.

Our approach of selecting utterances from a spontaneous conversation has shown how the complexity of the conversational data can be controlled. This means that we do not need a solution for synthesising all the speech phenomena in human conversation at once. Conversational speech can be used to approach conversational speech synthesis in a stepwise manner.

6.2.3 Blending Read Aloud and Conversational Speech

The challenge of blending is similar for both unit selection and HMM-based speech synthesis. It consists of utilising both conversational and read aloud speech data to synthesise speech that can convey an impression of a natural conversational style. This required the developed blending techniques to take into account the differences in language composition and general phonetic properties of the conversational and read aloud speech data (see section 3.5).

The particular engineering solutions of the unit selection and HMM-based frameworks had consequences for the details of the blending challenges as outlined in sections 2.3.2.2 and 2.3.3.6. In unit selection, the units to synthesise the discourse markers and filled pauses can be selected from discourse markers and filled pauses from a similar utterance position. Thus, the discourse markers and filled pauses maintain the natural phonetic properties of speech from a spontaneous conversation. The main challenge for unit selection lies in synthesising the propositional content of the utter-

ance. If we synthesise the propositional content with units from only neutrally read aloud speech, we are unlikely to have synthesised speech that conveys an impression of natural conversational speech. This is due to the fact that people can distinguish between natural read aloud and spontaneous speech in longer, utterance-size, stretches of speech (Blaauw, 1992). If we attempt to synthesise all the propositional content by selecting units from the conversational data, we are unlikely to synthesise speech without acoustic artefacts due to gaps in the phonetic coverage. Therefore, we designed the blending to select units from the conversational data when there existed appropriate units in the recorded conversational data, to synthesise speech with the phonetic properties of natural conversational speech. But, when appropriate conversational units were not available in the conversational data, then the units were selected from the read aloud data to maintain high quality synthesis without acoustic artefacts. The developed blending technique was described in section 4.3.3.

The HMM-based blending technique was described in section 4.2.5. In general, more data results in better quality synthetic speech (Aylett and Yamagishi, 2008). Building the HMM-based voice from the combined read aloud and conversational data would therefore result in better estimated statistical distributions of phonetic properties, than if we built HMM-based voices from just the conversational or read aloud data. However, uncontrolled use of both these data sources in HMM-based speech synthesis would likely result in acoustic artefacts, just as it did for unit selection (see section 5.4). The HMM-based blending therefore needed to allow boosting the phonetic coverage of our recorded conversational data by augmenting it with read aloud sentences, while still preserving the distinguishing phonetic properties in the two data sources. This was made by adding a speaking style context, spontaneous or read aloud, to the context-dependent phonemes. This speaking style context was then available during training to preserve any distinguishing characteristics between the conversational and read aloud data while allowing training speech models from both the data sources. The method was selected because it had previously been applied to preserve different “emotional” speaking styles in HMM-based speech synthesis (Yamagishi et al., 2005).

The blending allowed us to address the lack of coverage in spontaneous speech resources. Our analysis of phonetic properties of conversational and read aloud speech data suggested that the important aspects of blending were to smooth out any general acoustic differences between the data sources, while utilising the natural phonetic properties of the differences in language composition. The extent to which this could be made is discussed further in relation to the results from the perceptual experiments

in section 6.3.

6.2.4 Evaluating Conversational Speech Synthesis

Intuitively, the most suitable evaluation criterion for conversational speech synthesis is how natural the speech sounds in conversation. However, we have found little support in the literature that evaluating the quality of synthetic speech in a conversational discourse would be superior to evaluating isolated utterances. In natural speech, people can distinguish between whether someone is speaking spontaneously or is reading aloud from a script (Blaauw, 1992, 1994; Laan, 1997). These studies were all made on isolated utterances. Thus, there is no need for a conversational discourse in order for listeners to identify an utterance as coming from a spontaneous conversation. The added complexity of evaluating conversational characteristics in synthetic speech comes from the fact that the quality of synthetic utterances are likely to vary not only in their "spontaneity" or "conversational style", but also in how natural they sound.

Therefore, we evaluated these two aspects of our synthetic speech: naturalness and conversational style. Naturalness was selected as evaluation criterion to capture the extent to which the inclusion of speech from a spontaneous conversation and blending it with read aloud sentences impacted the acoustic artefacts and artificial quality of the synthetic speech. Conversational style was selected as an evaluation criterion to capture whether we could synthesise utterances typical for conversational speech with appropriate conversational characteristics. In addition to these two criteria we evaluated the pragmatic aspect of the synthetic speech. If we can synthesise conversational style speech, where discourse markers and filled pauses are wrapped around the propositional content, in a natural manner, does it affect the extent to which we can convey pragmatic information with the synthetic speech? To exemplify this, we evaluated to which extent we could alter the conveyed certainty about a proposition by altering the type of discourse markers and filled pauses that preceded the proposition.

Our evaluation criteria, naturalness, conversational style and pragmatic function, identify which aspects of the synthetic speech that needs to be improved. The main issue with our three evaluation criteria is that we cannot use just one of them to evaluate whether we have synthesised speech that conveys an impression of sounding like speech from a natural spontaneous conversation. Speech can be natural without sounding like speech coming from a spontaneous conversation, like the recordings of neutrally read aloud sentences conventionally used for speech synthesis. Synthetic speech

can convey an impression of sounding more spontaneous while still being distinctly less natural than conventional synthetic speech, as in the unit selection pilot experiment in section 4.3.2. Conveying a pragmatic function, e.g. a request "Pass me the salt!" or a greeting "Hi, how are you doing?", does not require the speech to sound either natural or conversational; in many cases the speech just needs to be intelligible. However, more natural deliveries of certain speech phenomena can convey certain pragmatic information more effectively, as was shown in the experiment in section 5.6.

Our evaluation criteria allowed us to establish to what extent the inclusion of conversational speech in synthetic voices affected the synthesis of conversational speech phenomena. The designs of our perceptual experiments and the formulations of our evaluation questions could most likely be refined. But, the results from the perceptual evaluations in chapter 4 and chapter 5 are conclusive. The inclusion of conversational speech data affected the naturalness of synthesising discourse markers and filled pauses in both unit selection and HMM-based speech synthesis. There was a strong correlation between naturalness and conversational style found in the experiment in section 5.2, but they were not the same. In all our experiments, listeners have readily identified the unit selection voice that contained speech from a conversation, irrespective of their naturalness, whereas the only time this happened for the HMM-based voices was when the "conversational" utterances were also perceived as more natural.

6.3 Summary of Results

The blending techniques and building of unit selection and HMM-based synthetic voices were described in chapter 4. In this section we will summarise the results from the phonetic and perceptual evaluations of the synthetic voices in chapter 4 and chapter 5.

6.3.1 Phonetic Analysis

The phonetic analysis of the synthetic speech from the HMM-based voices in section 4.2.6 showed that the HMM blending preserved phonetic properties of frequent conversational speech phenomena as well as a voice built from only conversational speech. But, there were also tendencies, e.g. in the speaking rate shown in figure 4.3, that the HMM blending did not preserve a distinction between the two speaking styles in the source data conversational or read aloud.

In the unit selection voices, the units were selected from either conversational or read aloud data. The unit selection blending method regulated the selection of units to avoid acoustic artefacts by selecting read aloud units when there was a gap in the phonetic coverage of the conversational speech data. This allowed the rest of the utterance to be synthesised using units with the natural phonetic properties of conversational speech. The analysis in section 4.3.6 showed that for a large set of in-domain text our unit selection blending approach selected the majority of units (62%) from the conversational data.

6.3.2 Perceptual Evaluations

Our hypothesis was that we could utilise conversational speech to add conversational characteristics to conventional synthetic voices without causing a negative impact on the naturalness. This was tested by evaluating two aspects of our synthetic voices: their naturalness and their conversational style.

As outlined in section 2.3.3.6, the main challenge with the HMM-based voices was to preserve an impression of a conversational style to listeners. The perceptual evaluation in section 5.2 therefore focused on investigating to what extent there was a difference in perceived style between our HMM-based voices. Two different groups of listeners were requested to evaluate either the naturalness or conversational style of utterances from the HMM-based voices. The tested HMM-based voices were built from a) conversational speech, b) read aloud speech or c) both read aloud and conversational speech (blended). The results showed that there were distinct differences in perceived naturalness. The inclusion of conversational speech made a clear positive impact on the quality of synthesised discourse markers and filled pauses, due to the phonetic content of the conversational speech data. But, we could not synthesise utterances with discourse markers and filled pauses that sounded as natural as more conventional material. The results from evaluating the conversational style of the HMM-based voices were less straightforward to interpret. The statistical analysis showed that the voice built from only conversational speech had a more conversational style than the voice built from only read aloud speech. Additionally, that analysis did not support that the blended voice preserved a distinction between the read aloud and conversational style of the source data. However, there was a strong correlation between the different groups' judgements of naturalness and conversational style, to the extent (Spearman's $\rho > 0.7$) where we doubt whether there was any meaningful difference related to con-

versational style, or if the results were related just to naturalness. This interpretation goes for the blended voice, which showed no difference in style, as well as the voices built from just conversational or read aloud data, which seemingly did show a difference in style. The conclusion drawn was that the inclusion of conversational speech data in HMM-based voices resulted in more natural-sounding conversational speech phenomena compared to a conventional voice. But, the synthesis of conversational speech phenomena was not natural enough for listeners to reliably identify characteristics associated with the differences between natural conversational and read aloud speech. The difference between "more natural-sounding conversational characteristics" and "more conversational style" is subtle but important. The statistical modelling in HMM-based synthesis results in a loss of the phonetic detail that allowed listeners in Blaauw (1992, 1994) to distinguish between natural spontaneous and read aloud speech.

The challenge of unit selection blending, as outlined in section 2.3.2.2, can be summarised as a matter of satisfying two conditions:

1. selecting enough conversational units to convey an impression of natural conversational speech

2. avoid introducing acoustic artefacts due to the lack of phonetic coverage.

To meet the first condition, the conversational speech needs to be accurately segmented and the speech phenomena in the data appropriately represented to allow synthesising them. These challenges were addressed in sections 3.4.5 and 3.3. To meet the second condition, we blended the conversational speech with read aloud speech. This blending took into account the differences in language composition of the two data sources, to determine when there was a lack of conversational coverage. Additionally, the blending itself may introduce acoustic artefacts. Therefore, the acoustic properties of the speech need to overlap to a sufficient degree to make the blending smooth and not stand out to the listeners. The results from the experiment in section 5.4 showed that our developed blending method, including a modified forced alignment and speaking rate adjustments, were required in order to not impair the synthetic speech quality when building voices from both the conversational and read aloud data. To evaluate whether we could utilise the blending method to add conversational characteristics to a conventional voice we designed the experiment in section 5.5. The experiment evaluated whether the blended voice could synthesise utterances typical for conversational speech in a manner that conveyed an impression to listeners of a conversational style.

The results showed that the conveyed conversational style was not due only to the content of the synthesised text, but due to the more appropriate synthesis of conversational speech phenomena. The qualitative analysis of the blending in section 5.5.2.1 suggested that some blending did not have a negative impact on quality, e.g. the read aloud units in one or two content words shown in bold face in the utterances:

uh you know but um, you know, I went there, and so uh, I was there for a few **weeks**.

long *story short it's garbage, my god, um, it is* **garb***age*

This is exactly what we need from the blending: enabling a smooth patch of the gaps in an utterance with read aloud units where the rest of the units are selected from a spontaneous conversation. But, there were limitations to the extent to which blending could be made. Utterances where stretches of more than a few words were selected from read aloud units were often still perceived by listeners as more conversational, but also less natural than speech from a conventional non-blended voice. Two examples of utterances where the blending was perceived as less natural are shown below. The long stretches of units selected from read aloud speech are shown in bold face:

uh then so, I **just wanna throw something**

uh you know, but as far as **getting** *out the* **theatres it has not done well**

The unit selection results support the findings in Blaauw (1992) where listeners made accurate distinctions between natural read aloud and spontaneous speech for segments of speech containing several words, but not for shorter, syllable-sized, segments.

We focused on evaluating to what extent we could achieve a natural and coherent conversational style in our blended voices. This was made to evaluate the general contribution on synthesising utterances with the structure and content typical for conversational speech, to show that we could integrate discourse markers and filled pauses with propositional content irrespective of the particular meaning of an utterance. To exemplify what we can communicate by synthesising natural conversational characteristics, we conducted the experiment in section 5.6. This experiment evaluated to what extent we could convey certainty or uncertainty about a topic by integrating discourse markers or filled pauses with propositional content. We designed two types of utterance initiation sequences that would potentially convey a difference in certainty about a topic:

- certainty: *oh yeah, three hundred dollars of sushi*

- uncertainty *um, you know uh, three hundred dollars of sushi*

These utterances were synthesised with the conventional unit selection voice and with the blended conversational voice. Natural utterances with a similar structure and content were also included in the evaluation. The results showed that all the natural and the synthetic voices conveyed a pragmatic contrast in the communicated degree of certainty between the two types of utterances. Neither synthetic voice communicated the degree of certainty as effectively as the natural speech. But, the blended conversational voice communicated the difference in certainty more effectively than the conventional voice. Hence, more natural-sounding conversational characteristics in synthetic voices results in more effective communication of pragmatic functions.

6.4 Conclusion

In this thesis we have shown how speech from a spontaneous conversation can be utilised to add more natural-sounding conversational characteristics to unit selection and HMM-based synthetic voices.

Our approach was to augment the conventionally used neutral read aloud sentences with speech from a spontaneous conversation. This blending approach was conceived out of the necessity to compensate for the lack of phonetic coverage in spontaneous speech resources. The approach was shown to be efficient in adding conversational characteristics to a conventional unit selection voice. The blended voice conveyed a general impression of a conversational style, as well as being more effective than the conventional voice in conveying a pragmatic contrast between certainty and uncertainty about a topic. The blending was most effective when only short sequences of read aloud segments needed to be mixed in with conversational units.

The blending method had a less positive impact in HMM-based speech synthesis. The results did not support that the blending preserved a perceptual distinction between the read aloud and conversational speaking styles in the source data. However, the strength of the HMM-based speech synthesis framework is to compensate for the lack of coverage. An HMM-based voice built from only conversational speech sounded more natural than a voice built from only conventional read aloud sentences, when the synthesised material contained conversational speech phenomena. The problem is that just sounding more natural when synthesising conversational material is insufficient.

The different evaluation criteria used in this thesis provide complementary information about the quality of the voices. It is important that the voices convey the in-

tended pragmatic information, like the degree of certainty about a topic, in the unit selection voices. But, pragmatic function is not sufficient. All the voices in our evaluation conveyed a pragmatic contrast due to the differing linguistic content of the utterances. To borrow the example from Campbell (2005): *Hi, how are you doing?*. This utterance is a greeting almost regardless of how badly it is synthesised, because it is the most likely pragmatic interpretation. To what extent we can avoid acoustic artefacts when synthesising this utterance, will be important for determining how natural it sounds. But, on top of these two basic criteria, there is a third that we evaluated as conversational style. Given two voices that can synthesise *Hi, how are you doing?* without acoustic artefacts, what determines the suitability for a believable character is to what extent the voices can synthesise the greeting in a natural conversational manner, i.e. convey an impression of conversational style.

It is more challenging to use speech from a spontaneous conversation than to use the conventional read aloud sentences to build synthetic voices. The motivation for using speech from a spontaneous conversation was that it provided a rich source of natural conversational speech phenomena. We selected utterances from a spontaneous conversation that contained a subset of these speech phenomena. This controlled set of conversational data could then be utilised to synthesise frequent conversational speech phenomena in unit selection and HMM-based synthetic voices. Thus, rather than the conventional method of designing prompts for a voice talent, natural spontaneous speech can be restricted to create a viable speech resource for building synthetic voices. In our opinion, such data is a better starting point for creating synthetic voices that can fulfil the vision in Loyall (1997) of interacting with your favourite movie character. If we use data-driven methods to create voices for such believable characters, perhaps from the movie itself, the data will be more expressive, contain a richer variety of speech phenomena, and have less controlled phonetic coverage than the conventionally used speech resources of neutrally read aloud sentences. These data properties are similar to the speech from a spontaneous conversation that we utilised in this thesis to build more conversational style voices. Therefore, we believe that learning to utilise such richer data sources is an important step towards creating synthetic voices that can express themselves in a natural conversational manner.

Appendix A

Test sentences

Utt. No.	Text
1	were you uh, serious when you were suggesting continuing the conversation or was that a subtle ploy to get me back into the uh, into recording studio
2	oh gosh, um, I'm not, to the perfect answer is I think, I could say I'm not sure and I was just in the wrong place at the right time, or is that the right place at the wrong time, or the wrong place at the wrong time, depending on your viewpoint, uh
3	no I mean I just
4	I've done a bit of recording work before
5	I like, I was gonna say I like speaking and I like, I like the sound of my own voice but not in a bad way, um
6	yeah, yeah, um I mean I've done some acting um
7	a little bit now and again nothing major I'm actually in a, I'm doing a couple of bits at the moment
8	to go back to that earlier conversation if we're reading a piece of text, then I think the text
9	if you read it in a casual way as if you're just chatting that may be fine for, if it's a, a part, if it's something to do with
10	you know a character saying something but if it's like, a speech or some piece of text which has, significance or has great moment then I think it's worth reading it in a particular way
11	and if you make it sound like bus conversation, then
12	you can get, good speeches and, good dialogue in films because that's one of the attitude not the attitudes that's one of the, um, uh what's the word I'm looking for here
13	it's still, well crafted or hopefully, is well crafted well constructed and comes across naturally but as a, a significant or interesting piece of text
14	trying to think of, particular, sorry does this, when I talk I'm turning my head a bit, is that making any difference to the
15	right okay
16	I'll try and uh restrict my movements a little bit, um
17	I was trying to think of some films where I thought there were good
18	I suppose the classic example of, the, joining of theatre and film, is when you take a film of a play
19	like for example Laurence Olivier in, the wartime Henry the fifth

Table A.1: Test sentences used for the HTS pilot experiment in section 4.2.3. The commas show where utterance internal pauses were located.

Utt. No.	Text
1	The more lot came with the HOUSE and the lower the price.
2	I just threw them on the side, INTENDING to transplant them, or throw them away or something.
3	My impression of it is that it has doubled in the last TEN years, and tripled in the last twenty.
4	They tried both soft CONVERSION and hard conversion.
5	My impression of it is that it has doubled in the last ten months and TRIPLED in the last twenty.
6	The MORE lot came with the house and the lower the price.
7	I just threw them on the side, INTENDING to transplant them, or throw them away or SOMETHING.
8	In that country, country women are in the BACKGROUND and the men are in the foreground.
9	They tried both SOFT conversion and hard conversion.
10	My impression of it is that it has DOUBLED in the last ten months and tripled in the last twenty.
11	They don't expand or contract when the WEATHER changes.
12	The more lot came with the car and the LOWER the price.
13	Sometimes the more YOU get the more you want too.
14	In that country, country women are in the background and the MEN are in the foreground.
15	I felt probably WORSE for them than for me.

Table A.2: Test sentences used for the HTS emphasis experiment in section 4.2.7. The word in upper case show which word was emphasised in an utterance.

Text
uh no, no well not, yet, um
so, um, but you have to live with yourself at the end of the day
but uh, uh I think it's an interesting enough story, uh so it's kind of a crime drama
yeah, X-men is cool, yeah
right, oh you you have to to transcribe all this
you know um boxing for me was more, uh it was far more challenging

Table A.3: Test sentences used for the HTS pilot experiment in section 4.2.7. The commas indicate where the utterance internal pauses were located.

Pair No.	Voice	Text
1	joh.16k.hts.spon	right, yeah that that could make you kind of a freak
	joh.16k.hts.read	you know um boxing for me was more, uh it was far more challenging
2	joh.16k.hts.spon	you know um boxing for me was more, uh it was far more challenging
	joh.16k.hts.read	uh no, no well not, yet, um
3	joh.16k.hts.spon	uh no, no well not, yet, um
	joh.16k.hts.read	yeah, X-men is cool, yeah
4	joh.16k.hts.spon	yeah, X-men is cool, yeah
	joh.16k.hts.read	right, oh you you have to to transcribe all this
5	joh.16k.hts.spon	right, oh you have to to transcribe all this
	joh.16k.hts.read	so let's see, but um, yeah, nothing exciting
6	joh.16k.hts.spon	so let's see, but um, yeah, nothing exciting
	joh.16k.hts.read	you know like when a, you go oh shit 'cause they didn't expect that
7	joh.16k.hts.spon	you know like when a, you go oh shit 'cause they didn't expect that
	joh.16k.hts.read	um, like a lot of people think I am in my late twenties
8	joh.16k.hts.spon	um, like a lot of people think I am in my late twenties
	joh.16k.hts.read	so, it's uh, yeah, mid-life crisis got it's gonna hit eventually, pretty quickly so
9	joh.16k.hts.spon	so, it's uh, yeah, mid-life crisis got it's gonna hit eventually, pretty quickly so
	joh.16k.hts.read	yeah, I could give a shit less um I'm just happy to get a meal
10	joh.16k.hts.spon	yeah, I could give a shit less um I'm just happy to get a meal
	joh.16k.hts.read	um, but even that like, I can give a shit less, you know what I mean
11	joh.16k.hts.spon	um, but even that like, I can give a shit less, you know what I mean
	joh.16k.hts.read	oh yeah you don't want that to happen
12	joh.16k.hts.spon	oh yeah you don't want that to happen
	joh.16k.hts.read	well we quit I mean you know the movie ended
13	joh.16k.hts.spon	well we quit I mean you know the movie ended
	joh.16k.hts.read	yeah I just fill in my schedule so it's uh
14	joh.16k.hts.spon	yeah I just fill in my schedule so it's uh
	joh.16k.hts.read	no I have well you know I tried once when I was a kid
15	joh.16k.hts.spon	no I have well you know I tried once when I was a kid
	joh.16k.hts.read	right, yeah that that could make you kind of a freak

Table A.4: Test sentence pairs for the evaluation in section 5.2. Commas indicate where utterance internal silences were located.

Pair No.	Voice	Text
1	joh.16k.hts.blendspon	right, yeah that that could make you kind of a freak
	joh.16k.hts.blendread	boxing for me was more, it was far more challenging
2	joh.16k.hts.blendspon	you know um boxing for me was more, uh it was far more challenging
	joh.16k.hts.blendread	well not yet
3	joh.16k.hts.blendspon	uh no, no well not, yet, um
	joh.16k.hts.blendread	x-men is cool
4	joh.16k.hts.blendspon	yeah, X-men is cool, yeah
	joh.16k.hts.blendread	you have to transcribe all this
5	joh.16k.hts.blendspon	right, oh you have to to transcribe all this
	joh.16k.hts.blendread	let's see, but nothing exciting
6	joh.16k.hts.blendspon	so let's see, but um, yeah, nothing exciting
	joh.16k.hts.blendread	when you go, shit 'cause they didn't expect that
7	joh.16k.hts.blendspon	you know like when a, you go oh shit 'cause they didn't expect that
	joh.16k.hts.blendread	a lot of people think I am in my late twenties
8	joh.16k.hts.blendspon	um, like a lot of people think I am in my late twenties
	joh.16k.hts.blendread	mid-life crisis, it's gonna hit eventually, pretty quickly
9	joh.16k.hts.blendspon	so, it's uh, yeah, mid-life crisis got it's gonna hit eventually, pretty quickly so
	joh.16k.hts.blendread	I could give a shit less, I'm just happy to get a meal
10	joh.16k.hts.blendspon	yeah, I could give a shit less um I'm just happy to get a meal
	joh.16k.hts.blendread	but even that I can give a shit less
11	joh.16k.hts.blendspon	um, but even that like, I can give a shit less, you know what I mean
	joh.16k.hts.blendread	you don't want that to happen
12	joh.16k.hts.blendspon	oh yeah you don't want that to happen
	joh.16k.hts.blendread	we quit, the movie ended
13	joh.16k.hts.blendspon	well we quit I mean you know the movie ended
	joh.16k.hts.blendread	I just fill in my schedule
14	joh.16k.hts.blendspon	yeah I just fill in my schedule so it's uh
	joh.16k.hts.blendread	I have, I tried once when I was a kid
15	joh.16k.hts.blendspon	no I have well you know I tried once when I was a kid
	joh.16k.hts.blendread	that could make you kind of a freak

Table A.5: Test sentence pairs for the evaluation in section 5.2. Commas indicate where utterance internal silences were located.

Utt. No.	Text
1	yeah, I prefer Glasgow there's a lot more variety as well
2	and uh, it's a bit less
3	there's some of the clubs in Edinburgh where there is a very particular crowd that goes and it's quite cliquey and
4	you would never go there but, well some people do, obviously
5	um, but yeah I think I prefer Glasgow
6	yeah definitely, plenty of that going on
7	although I was really lucky my, my supervisor was great, the only, the only thing I could say against her was the fact she's a Hibs supporter which uh
8	definitely counts against her
9	yeah maybe, they well they
10	yeah they probably don't speak like the queen
11	although there are a few people that, come from around Edinburgh and, you think that they come from the south of England and then they say no no I lived here my, my whole life it's just cause they went to really nice schools and stuff
12	yeah, yeah I went um, when I was coming to uni I, um, went down to Cambridge for an interview
13	and uh, I met this girl who was Welsh
14	but, I didn't realise that for, the majority of the conversation I had with her she was she'd been studying there for three years or something and they'd got her to come in and help the, the kind of new recruits to settle in and stuff
15	and she just sounded like she was from somewhere, you know Kent or somewhere like, really nice in the south east of England
16	and then her mom phoned and, when she spoke to her mom she went back into her own Welsh accent which was so strong, and much nicer
17	and, I was just like I, I would hate, to, to change like that and not be true, anymore to, to where I come from, to, it was
18	it just seemed like she was ashamed of her own accent and you know she didn't fit in and, people might not always have understood her first time round so she just adapted but
19	I think it, I thought it was awful, it was horrible

Table A.6: Test sentences used for the unit selection pilot experiment in section 4.3.2. The commas show where utterance internal pauses were located.

Utt. No.	News Sentences
1	Soldiers have lived a precarious existence within the posts, using state of the art listening devices and long range cameras to maintain round the clock surveillance.
2	In peshawar, the capital of the north west frontier province which lies on the border with Afghanistan, the influx of refugees means that there are now more Afghans than local people
3	He had a showbusiness lifestyle, driving a porsche nine eleven and living in a luxurious home in the south west of London.
4	The european union yesterday agreed a ban on four antibiotics used by farmers to fatten livestock, amid fears that the practice reduces the effectiveness of life saving drugs
5	Far from narrowing the definition of Scottish citizenship we want a wider and more inclusive definition of citizenship that the definition of UK citizenship is at the moment.
6	The pope's comments came as he began a five day trip to the muslim nation of Azerbaijan, before heading to Bulgaria.
7	From early on, the administration has argued that Iraq's cooperation was insincere, and that Saddam was toying with the inspectors while earning himself precious time
8	The only people trying to keep an accurate track of casualties is a network of soldiers' mothers' associations scattered from Murmansk to the Black Sea.
9	Safeway is half way through a two year investment plan, spending one hundred million pounds and creating two thousand new jobs in Scotland.
10	Both euro MP's are spearheading a drive for better information to be given on the dangers of implants before and after cosmetic surgery.

Table A.7: Test sentences used for experiment in section 5.4.

Utt. No.	Conversational Sentences
1	*You know um boxing for me was more, uh it was far more* **challenging**
2	*Uh no, no* **well** *not, yet, um*
3	*Yeah,* **X-men** *is cool, yeah*
4	*So* **let's see,** *but um, yeah, nothing* **exciting**
5	*Yeah, I could give a shit* **less** *um Im just* **happy** *to get a* meal
6	*Um, but even that* **like,** *I can give a shit* **less,** *you know what I mean*
7	*Oh yeah you don't want that to* **happen**
8	*Well* **we quit** *I mean you know the movie* **ended**
9	*Yeah I just* **fill** *in my schedule so its uh*
10	*No I have well you know I tried* **once** *when I was a kid*

Table A.8: Test sentences used for experiment in section 5.4. The italic text shows where the unit selection was biased towards selecting units from the conversational data. The bold faced text shows where selection of read aloud units was likely due to the lack of coverage of conversational units.

Utt. No.	Test sentences
0	I wasn't too embarrassed to say that's disgusting. *You know uh, I wasn't* **too embarrassed** *to say* **that's disgusting.**
1	I'm about to kill him and I get arrested and so forth. *You know uh, I'm about to kill him and I get* **arrested** *and so forth uh.*
2	There's the whole question between, babying and nurturing your child. *Uh, there's the whole* **question between, babying and nurturing your child.**
3	No, that would be really cool. *Yeah no, that would be really cool.*
4	I just don't do that kind of thing. *Yeah, I just don't do that kind of thing um.*
5	That's the worst part about being an actor. *Yeah so, that's the worst part about being an actor.*
6	Then I just wanna throw something. *Uh then so, I just* **wanna throw something.**
7	That helped with our domestic sales and, internationally it's done well. *Like, that* **helped with our domestic sales and, internationally,** *you know, it's done* **well.**
8	I pull myself about here and I've got a website with my name on it. *Yeah, I pull* **myself about** *here and I've got a* **website** *with my name on it.*
9	But as far as getting out the theatres it has not done well. *Uh you know, but as far as* **getting** *out the* **theatres it has not done well.**
10	But, I went there, and I was there for a few weeks. *Uh you know but um, you know,* **I went there,** *and so uh,* *I was there for a few* **weeks.**
11	It's a different character for me. *Uh it's um, a different* **character** *for me.*
12	Long story short it's garbage, my god it is garbage. *Long story short it's garbage, my god, um, it is* **garbage.**
13	It's all up to you guys to make me, sound good or bad or whatever. *Yeah so, it's all up to you guys to make me, yeah,* **sound good or bad** *or whatever.*
14	Apparently I know way too much about the sex stuff here in America. **Apparently,** *yeah, I know way too much* **about,** *like, the* **sex stuff** *here in* **America.**

Table A.9: Test sentences used for experiment in section 5.5. The italic text shows where the unit selection was biased towards selecting units from the conversational data. The bold faced text shows where selection of read aloud units was likely due to the lack of coverage of conversational units.

Utt. No.	CERT Sentences
1	*oh yeah, a* **list of conversations**
2	*yeah, a* **vast majority of the members**
3	*right yeah, about* **two years ago**
4	*oh yeah,* **three hundred dollars of sushi**
5	*yeah,* **fourteen matches in eight years**
6	*right yeah,* **except for take off and landing**
7	*oh yeah,* **make some decent money**
Natural	yeah, in Monday they're buying
Natural	yeah, no I can talk without needing a break
Utt. No.	UNCERT Sentences
1	*um, you know uh, a* **list of conversations**
2	*you know um, uh, a* **vast majority of the members**
3	*well, you know um, about* **two years ago**
4	*um, you know uh,* **three hundred dollars of sushi**
5	*you know um, uh,* **fourteen matches in eight years**
6	*well, you know um,* **except for take off and landing**
7	*um, you know uh,* **make some decent money**
Natural	um, no I uh, uh I moved up for, acting
Natural	I just saw um, uh um, a version of a film that I did, um, in Thai

Table A.10: Test sentences used for experiment in section 5.6. The italic text shows where the unit selection was biased towards selecting units from the conversational data. The bold faced text shows where selection of read aloud units was likely due to the lack of coverage of conversational units. The table also shows the transcripts of the natural reference samples used in the evaluation.

Bibliography

Adell, J., Bonafonte, A., and Escudero, D. (2006). Disfluent speech analysis and synthesis: A preliminary approach. In *Speech Prosody*, Dresden, Germany.

Adell, J., Bonafonte, A., and Escudero, D. (2007a). Filled pauses in speech synthesis: Towards conversational speech. In *10th International Conference TSD 2007*, pages 358–365, Pilsen, Czech Republic.

Adell, J., Bonafonte, A., and Escudero, D. (2007b). Statistical analysis of filled pauses' rhythm for disfluent speech. In *SSW6*, pages 223–227, Bonn, Germany.

Adell, J., Bonafonte, A., and Escudero-Mancebo, D. (2008). On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms. In *Interspeech*, pages 2278–2281, Brisbane, Australia.

Adell, J., Bonafonte, A., and Escudero-Mancebo, D. (2010). Modelling filled pauses prosody to synthesise disfluent speech. In *Speech Prosody*, volume 100624, pages 1–4, Chicago, U.S.A.

Andersson, S., Badino, L., Watts, O., and Aylett, M. (2008). The CSTR/CereProc Blizzard entry 2008: The inconvenient data. In *The Blizzard Challenge*, Brisbane, Australia.

Andersson, S., Georgila, K., Traum, D., Aylett, M., and Clark, R. (2010a). Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Speech Prosody*, volume 100116, pages 1–4, Chicago, U.S.A.

Andersson, S., Yamagishi, J., and Clark, R. (2010b). Utilising spontaneous conversational speech in HMM-based speech synthesis. In *SSW7*, pages 173–178, Kyoto, Japan.

Andersson, S., Yamagishi, J., and Clark, R. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication*, 54(2):175–188.

Arnold, J., Hudson Kam, C., and Tanenhaus, M. (2007). If you say *Thee uh* you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33(5):914–930.

Aylett, M. and Pidcock, C. (2007). The CereVoice characterful speech synthesiser SDK. In *AISB'07*, pages 174–178, Newcastle Upon Tyne, U.K.

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Aylett, M. and Yamagishi, J. (2008). Combining statistical parametric speech synthesis and unit-selection for automatic voice cloning. In *LangTech*, Rome, Italy.

Badino, L. (2010). *Identifying Prosodic Prominence Patterns for English Text-to-Speech Synthesis*. PhD thesis, University of Edinburgh, Edinburgh, U.K.

Badino, L., Andersson, S., Yamagishi, J., and Clark, R. (2009). Identification of contrast and its emphatic realisation in HMM based speech synthesis. In *Interspeech*, pages 520–523, Brighton, U.K.

Bell, A., Jurafsky, D., Fossler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

Benus, S., Gravano, A., and Hirschberg, J. (2007). The prosody of backchannels in American English. In *ICPhS*, pages 1065–1068, Saarbrücken, Germany.

Blaauw, E. (1992). Phonetic differences between read and spontaneous speech. In *ICSLP*, pages 751–754, Banff, Canada.

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, 14(4):359–375.

Black, A. and Tokuda, K. (2005). The Blizzard challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In *Interspeech 2005*, pages 77–80, Lisbon, Portugal.

Brennan, S. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34(3):383–398.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Chengyu Fang, A., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D. (2010). Towards an ISO standard for dialogue act annotation. In *7th International Conference on Language Resources and Evaluation*, pages 2548–2555, Valetta, Malta.

Cabral, J., Renals, S., Richmond, K., and Yamagishi, J. (2008). Glottal spectral separation for parametric speech synthesis. In *Interspeech*, pages 1829–1832, Brisbane, Australia.

Cadic, D. and Segalen, L. (2008). Paralinguistic elements in speech synthesis. In *Interspeech*, pages 1861–1864, Brisbane, Australia.

Campbell, N. (2005). Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE Transactions on Information and Systems*, E88-D(3):376–383.

Campbell, N. (2006). On the structure of spoken language. In *Speech Prosody*, Dresden, Germany.

Campbell, N. (2007). Towards conversational speech synthesis; lessons learned from the expressive speech processing project. In *SSW6*, pages 22–27, Bonn, Germany.

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

Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, U.K.

Clark, H. and Fox Tree, J. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Clark, R., Richmond, K., and King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4):317–330.

Corley, M., MacGregor, L., and Donaldson, D. (2007). It's the way that you, er, say it: hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.

Fox Tree, J. (2001). Listener's uses of um and uh in speech comprehension. *Memory & Cognition*, 29(2):320–326.

Fuller, J. (2003). The influence of speaker roles on discourse marker use. *Journal of Pragmatics*, 35(1):23–45.

Goffman, E. (1967). *Interaction Ritual - Essays on Face-to-Face Behavior*. Anchor Books, New York, U.S.A.

Gravano, A., Benus, S., Chavez, H., Hirschberg, J., and Wilcox, L. (2007). On the role of context and prosody in the interpretation of "okay". In *ACL*, pages 800–807, Prague, Czech Republic.

Grosz, B. and Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Gustafsson, K. and Sjölander, K. (2004). Voice creation for conversational fairy-tale characters. In *5th ISCA Speech Synthesis Workshop*, pages 145–150, Pittsburgh, U.S.A.

Hanson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101(1):466–481.

Hanson, H. and Chuang, E. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America*, 106(2):1064–1077.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hockey, B. (1993). Prosody and the role of *Okay* and *Uh-huh* in discourse. In *Proc. of ESCOL*, pages 128–136, Columbus, U.S.A.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP*, pages 373–376, Atlanta, U.S.A.

Jakobovits, L. and Miron, M., editors (1967). *Readings in the Psychology of Language*, chapter Hesitation Phenomena in Spontaneous English Speech. Prentice Hall, New Jersey, U.S.A.

Johnson, K. (2004). Massive reduction in conversational American English. In *Proc. of the 1st session of the 10th international symposium on spontaneous speech: data and analysis*, pages 29–54, Tokyo, Japan.

Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120, Montreal, Canada.

Karaiskos, V., King, S., Clark, R., and Mayo, C. (2008). The Blizzard challenge 2008. In *The Blizzard Challenge*, Brisbane, Australia.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. 2nd MAVEBA*, Firenze, Italy.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207.

King, S. and Karaiskos, V. (2009). The Blizzard challenge 2009. In *The Blizzard Challenge*, Edinburgh, U.K.

King, S. and Karaiskos, V. (2010). The Blizzard challenge 2010. In *The Blizzard Challenge*, Kyoto, Japan.

Kominek, J. and Black, A. (2004). The CMU Arctic speech databases. In *5th ISCA Speech Synthesis Workshop*, pages 223–224, Pittsburgh, U.S.A.

Laan, G. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22(1):43–65.

Ladefoged, P. (2006). *A Course in Phonetics*. Thomson Wadsworth, Boston, U.S.A.

Lasarcyk, E. and Wollerman, C. (2010). Do prosodic cues influence uncertainty perception in articulatory speech synthesis? In *SSW7*, pages 230–235, Kyoto, Japan.

Lee, C.-H., Wu, C.-H., and Guo, J.-C. (2010). Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation. In *ICASSP*, pages 4826–4829, Dallas, U.S.A.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Lickley, R. (1994). *Detecting Disfluency in Spontaneous Speech*. PhD thesis, University of Edinburgh, Edinburgh, U.K.

Lickley, R. (1995). Missing disfluencies. In *ICPhS*, volume 4, pages 192–195, Stockholm, Sweden.

Lickley, R. (2001). Dialogue moves and disfluency rates. In *Proc. of DiSS'01*, pages 93–96, Edinburgh, U.K.

Local, J. (2007). Phonetic detail and the organisation of talk-in-interaction. In *ICPhS*, pages 1–10, Saarbrücken, Germany.

Local, J. and Walker, G. (2005). Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica*, 62(2-4):120–130.

Loyall, B. (1997). *Believable Agents: Building Interactive Personalities*. PhD thesis, Carnegie Mellon University, Pittsburgh, U.S.A.

Maclay, H. and Osgood, C. (1959). Hesitation phenomena in spontaneous english speech. *Word*, 15.

Mayo, C., Clark, R., and King, S. (2005). Multidimensional scaling of listener responses to synthetic speech. In *Interspeech*, pages 1725–1728, Lisbon, Portugal.

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer, Speech and Language*, 22(2):171–184.

Newell, C. (2009). *Place, Authenticity, and Time: a Framework for Liveness in Synthetic Speech*. PhD thesis, The University of York, York, U.K.

O'Connel, D. and Kowal, S. (2005). *Uh* and *Um* revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34(6):555–576.

Odell, J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, Cambridge, U.K.

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. In *ICASSP*, pages 521–524, San Fransisco, U.S.A.

Parviainen, O. (2012). SoundTouch. http://www.surina.net/soundtouch/soundstretch.html.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. (2008). HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In *Interspeech*, pages 1881–1884, Brisbane, Australia.

Romportl, J., Zovato, E., Santos, R., Ircing, P., Relano Gil, J., and Danieli, M. (2010). Application of expressive TTS synthesis in an advanced ECA system. In *SSW7*, pages 120–125, Kyoto, Japan.

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, 50(4):696–735.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge University Press, Cambridge, U.K.

Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., and Schuller, B. (2008). Towards responsive sensitive artificial listeners. In *Fourth International Workshop on Human-Computer Conversation*, Bellagio, Italy.

Searle, J. (1969). *Speech Acts: an Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, U.K.

Shriberg, E. (1996). Disfluencies in Switchboard. In *ICSLP*, Philadelphia, U.S.A.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *ICPhS*, pages 619–622, San Fransisco, U.S.A.

Shriberg, E. and Lickley, R. (1993). Intonation of clause-internal filled pauses. *Phonetica*, 50(3):172–179.

Shriberg, E. and Stolcke, A. (1996). Word predictability after hesitations: A corpus-based study. In *ICSLP*, pages 1868–1871, Philadelphia, U.S.A.

Silén, H., Helander, E., Nurminen, J., and Gabbouj, M. (2009). Parametrization of vocal fry in HMM-based speech synthesis. In *Interspeech*, pages 1775–1778, Brighton, U.K.

Strom, V., Clark, R., and King, S. (2006). Expressive prosody for unit-selection speech synthesis. In *Interspeech*, pages 1522–1525, Pittsburgh, U.S.A.

Strom, V., Nenkova, A., Clark, R., Vazquez-Alvarez, Y., Brenier, J., King, S., and Jurafsky, D. (2007). Modelling prominence and emphasis improves unit-selection synthesis. In *Interspeech*, pages 1282–1285, Antwerp, Belgium.

Sundaram, S. and Narayanan, S. (2002). Spoken language synthesis: Experiments in synthesis of spontaneous dialogues. In *2002 IEEE Speech Synthesis Workshop*, pages 203–206, Santa Monica CA, U.S.A.

Sundaram, S. and Narayanan, S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Eurospeech*, pages 1221–1224, Geneva, Switzerland.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2006). A style adaptation technique for speech synthesis using HSMM and suprasegmental features. *IEICE Transactions on Information and Systems*, E89-D(3):1092–1099.

Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, E90-D(5):816–824.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modelling. In *ICASSP*, pages 229–231, Phoenix, U.S.A.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *ICASSP*, pages 1315–1318, Istanbul, Turkey.

Tokuda, K., Zen, H., and Black, A. (2002). An HMM-based speech synthesis system applied to English. In *2002 IEEE Speech Synthesis Workshop*, pages 227–230, Santa Monica CA, U.S.A.

Traum, D., Swartout, W., Gratch, J., and Marsella, S. (2008). A virtual human dialogue model for non-team interaction. In Dybkjaer, L. and Minker, W., editors, *Recent Trends in Discourse and Dialogue*, pages 45–67, Antwerp, Belgium. Springer.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14(1):129–182.

Werner, S., Wolff, M., and Hoffmann, R. (2006). Pronunciation variant selection for spontaneous speech synthesis - listening effort as a quality parameter. In *ICASSP*, pages 857–860, Tolouse, France.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):66–83.

Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2005). Acoustic modelling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, E88-D(3):502–509.

Yamagishi, J., Tachibana, M., Masuko, T., and Kobayashi, T. (2004). Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. In *ICASSP*, volume 1, pages 5–8, Quebec, Canada.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.-J., Tokuda, K., Karhila, R., and Kurimo, M. (2010). Thousands of voices for HMM-based speech synthesis - analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):984–1004.

Yamagishi, J., Zen, H., Toda, T., and Tokuda, K. (2007). Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard challenge 2007. In *The Blizzard Challenge*, Bonn, Germany.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard challenge. In *The Blizzard Challenge 2008*, Brisbane, Australia.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book*. Cambridge University Engineering Department.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Transactions on Information and Systems*, E90-D(1):325–333.

Zen, H., Tokuda, K., and Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zen, H., Tokuda, K., and Kitamura, T. (2004a). An introduction of trajectory model into HMM-based speech synthesis. In *5th ISCA Speech Synthesis Workshop*, pages 191–196, Pittsburgh, U.S.A.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004b). Hidden semi-Markov model based speech synthesis. In *Interspeech*, pages 1393–1396, Jeju Island, South Korea.