Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis

Sebastian Andersson, Junichi Yamagishi, and Robert A.J. Clark. Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. Speech Communication, 54(2):175-188, 2012.

Synthesis and Evaluation of Conversational Characteristics in HMM-based Speech Synthesis

Sebastian Andersson*, Junichi Yamagishi, Robert A. J. Clark

The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, UK

Abstract

Spontaneous conversational speech has many characteristics that are currently not modelled well by HMM-based speech synthesis and in order to build synthetic voices that can give an impression of someone partaking in a conversation, we need to utilise data that exhibits more of the speech phenomena associated with conversations than the more generally used carefully read aloud sentences. In this paper we show that synthetic voices built with HMM-based speech synthesis techniques from conversational speech data, preserved segmental and prosodic characteristics of frequent conversational speech phenomena. An analysis of an evaluation investigating the perception of quality and speaking style of HMM-based voices confirms that speech with conversational characteristics are instrumental for listeners to perceive successful integration of conversational speech phenomena in synthetic speech. The achieved synthetic speech quality provides an encouraging start for the continued use of conversational speech in HMM-based speech synthesis.

Keywords: speech synthesis, HMM, conversation, spontaneous speech, filled pauses, discourse marker

1. Introduction

Unit selection and HMM-based synthetic voices can synthesise neutral read aloud speech quite well, both in terms of intelligibility and naturalness (King and Karaiskos, 2009). For many applications, e.g. in vehicle GPS systems, an intelligible read aloud speaking style is sufficient to provide a user with relevant information. But other applications, e.g. believable virtual characters (Traum et al., 2008), embodied conversational agents (Romportl et al., 2010) or speech-to-speech translation (Wahlster, 2000), require synthetic voices with more conversational characteristics that can synthesise turn-taking behaviour, provide backchannels and express agreement, disagreement, hesitation, et cetera.

The fundamental concept in both unit selection and HMM-based speech synthesis is the ability to utilise recordings of natural speech directly, and build synthetic voices that preserve the segmental and prosodic properties of the speech and the speaker in the recordings. To be able to synthesise the segmental and prosodic properties of various speech phenomena from recordings of speech, the recordings must contain examples of these speech phenomena. This has been demonstrated for HMM-based speech synthesis by Yamagishi et al. (2005) where recordings portraying emotional speaking styles were shown to be instrumental in listeners being able to perceive such styles in synthetic speech. Badino et al. (2009) demonstrated that recordings incorporating emphatic accents similarly improves the quality of emphatic accents in the resulting synthesis.

Whereas it may be theoretically possible to synthesise any style of speech from a corpora of neutrally read sentences, for example by manipulating the speech at the frame level, our current understanding of the relationship between the intention that a speaker wishes to convey and the resulting speech signal is such that doing this well, particularly in a conversational context, is extremely difficult.

As our current speech synthesis techniques are specifically formulated to reproduce the speech characteristics found in the data that they use, it is appropriate to attempt to use richer data and capture more of the natural characteristics found there. More specifically, in order to build synthetic voices that are better suited to generating realistic conversational speech, we should be building these voices from data that exhibits more of the characteristics we associate with conversational speech.

Scientific questions and problems arising for this goal include how to acquire spontaneous conversational speech in the high quality conditions that suit speech synthesis while maintaining the conversational speech phenomena that distinguish this speech from read speech, and how to best model these phenomena within our statistical frameworks. If we can solve these issues, they will provide a stepping-stone to being able to use the wealth of emerging informal speech data resources such as pod-casts and

*Corresponding author. TEL:+44-131-651-3284

Email addresses: J.S.Andersson@sms.ed.ac.uk (Sebastian Andersson), jyamagis@inf.ed.ac.uk (Junichi Yamagishi), robert@cstr.ed.ac.uk (Robert A. J. Clark)

speech available from Internet sites such as YouTube[1].

As the first step towards this goal, we utilised carefully selected speech from a spontaneous conversation to build synthetic voices with HMM-based speech synthesis techniques (Andersson et al., 2010b). These "conversational" voices were contrasted with HMM-based voices built from a more conventional data source of neutrally read aloud sentences in a perceptual evaluation designed to compare the quality and speaking style of the voices. The evaluation in the previous publication showed an interesting tendency of the listeners' perception: The "conversational" voice was perceived as being more natural than the "read aloud" voice, and was perceived as having a more conversational speaking style when the compared synthetic utterances contained certain lexical items that are frequent in conversations. But, when these conversational characteristics were removed from the sentences synthesised with the "read aloud" voice, the "conversational" voice was perceived as less natural and was no longer perceived as having a more conversational speaking style.

Therefore, in this paper we present the details of our speech synthesis systems built from spontaneous conversational speech and make a comparative analysis of segmental and prosodic properties of the natural and synthetic voices, together with an analysis of the perceptual evaluation of naturalness and conversational speaking style, to investigate why the presence or absence of a few words had such an impact on perceived speech quality.

The rest of the paper is organised as follows: Section 2 provides a background of previous research on conversational speech phenomena and synthetic speech with conversational characteristics. Section 3 gives an overview of the HMM-based speech synthesis system used to build our synthetic voices. Sections 4 and 5 describe the conversational and read aloud speech data. Sections 6 and 7 describe the building of synthetic voices and the phonetic and perceptual evaluations of the synthetic speech. Section 8 contains a final discussion and conclusions.

2. Background

2.1. Related work

Previous research synthesising speech with spontaneous or conversational characteristics has mainly been achieved with techniques other than HMM-based speech synthesis, e.g. unit selection speech synthesis (Cadic and Segalen, 2008; Andersson et al., 2010a; Adell et al., 2010), limited domain unit selection synthesis (Sundaram and Narayanan, 2002; Gustafsson and Sjölander, 2004), phrase level selection from a very large corpus (Campbell, 2007), or articulatory speech synthesis (Lasarcyk and Wollerman, 2010). The approach in Lee et al. (2010) did use HMM-based speech synthesis to model pronunciation variation of spontaneous speech, but has a different focus to that of dealing with the language differences in content and structure between read aloud and conversational speech.

2.2. Conversational Speech Phenomena

An inclusive definition of the differences in language structure and content between read aloud and conversational speech as being "wrappers" around propositional content was given in Campbell (2006). An example from our data is given below with the wrappers in italics and the propositional content in bold face:

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

Based on previous research regarding the phonetic and discourse properties of the wrapper category, we further divided wrappers into filled pauses, discourse markers and backchannels:

Filled pauses are generally regarded as a hesitation phenomena. The transliteration of English filled pauses differs slightly within the literature, but we will use *um* and *uh*. Filled pauses are word-like (Clark and Fox Tree, 2002), but their specific phonetic properties distinguishes them from other words in terms of vowel quality, F0 and on average a much longer vowel duration (O'Shaughnessy, 1992; Shriberg, 1999). As a hesitation phenomena they are often associated with a prolongation of at least the preceding syllable (Adell et al., 2008).

Discourse markers is one of many terms that have been used to refer to similar sets of words and expression that are used to regulate the flow of the conversation, rather than communicate propositional content (Schiffrin, 1987). Although discourse markers consist of lexical forms that exist also in the read aloud sentences, e.g. *okay*, *so*, *because* or *you know*, the phonetic properties of these words are different when used as discourse markers than when used in other functions (Schiffrin, 1987; Gravano et al., 2007). A frequent discourse marker in conversation is *yeah*, often used to signal agreement in the beginning of a turn (Jurafsky et al., 1998). It is also worth noting that Jurafsky et al. (1998) treat *yeah*, *oh yeah*, and *well yeah* as separate tokens that can share the same discourse function, e.g. agreement or backchannel.

Backchannels are signals that the listener is involved in the dialogue, but does not want to take the turn from the speaker (Gravano et al., 2007). They often have the same lexical realisation as discourse markers, e.g. *yeah* or *okay*. Although phonetic features, such as pitch slope, were different in manually classified backchannels than when used in other functions, an important classification cue was that backchannels were often isolated from other speech by the same speaker (Jurafsky et al., 1998; Gravano et al., 2007).

3. HMM-based Speech Synthesis

All the synthetic voices described in this paper were built with the speaker dependent HMM-based speech syn-

[1] http://www.youtube.com

thesis system (HTS) (Zen et al., 2007). The text analysis and generation of context dependent phonemes are not part of the standard HTS system but were added by us, in conjunction with using the CereVoice system (Aylett and Pidcock, 2007). The only difference between the voices described here and standard HTS voices is the speaking style of the data and the additional blending of speaking styles mentioned in Section 6.1. An overview of the acoustic feature extraction, training of HMM-based models, and generating synthetic speech is given below:

1. **Acoustic Feature Extraction:** Spectral and excitation parameters are extracted from the acoustic speech signal as STRAIGHT (Kawahara et al., 1999, 2001) mel-cepstrals, aperiodicity and $\log F0$ parameters.
2. **HMM Training:** The acoustic parameters together with the context dependent phoneme descriptions are jointly trained in an integrated HMM-based statistical framework to estimate Gaussian distributions of excitation ($\log F0$ and aperiodicity), spectral (STRAIGHT mel-cepstrals) and duration parameters for the context dependent phonemes.
3. **HMM Clustering:** Due to the large number of context combinations there are generally only a few instances of each combination and many combinations are not present in the training data. To reliably estimate statistical parameters for context combinations the data is shared between states in the HMMs through decision tree-based context clustering (Odell, 1995). The resulting clustered trees also enable dealing with unseen context combinations at the synthesis stage. Trees are constructed separately for mel-cepstrals, aperiodicity, $\log F0$ and duration.
4. **Speech Generation:** At the synthesis stage an input text sentence is converted into a context dependent phoneme sequence. Speech (spectral, excitation and duration) parameters are then generated from the corresponding trained HMMs and rendered into a speech signal through the STRAIGHT mel-cepstral vocoder with mixed excitation.

4. Spontaneous Conversational Speech Data

4.1. Recording

The spontaneous conversational speech data, used to build the synthetic voices in this paper, was recorded for the purpose of use in speech synthesis. The speech data consisted of manually transcribed and selected utterances, from a total of approximately seven hours of studio recorded conversation between the first author of this paper and an American male voice talent in his late thirties from Texas (Andersson et al., 2010a).

The voice talent was positioned inside the recording booth, and the author was positioned outside it. The speech from the voice talent and the author were recorded on separate channels, and the communication was made via microphones and headphones, but they had eye-contact through a window. The conversation was unconstrained, but mainly focused around the voice talent's work as an actor and his interests in films and martial arts.

4.2. Transcription and Selection

The manual transcription and selection of speech from the conversation were conducted by the first author to obtain speech data that would provide a sensible starting point for state-of-the-art statistical speech synthesis techniques.

First the speech of the voice talent was transcribed orthographically and aligned at the utterance level. The motivation for an orthographic transcription was that it is the description level generally assumed in text-to-speech synthesis, and it provides a token level suitable for subsequent manual or automatic processing.

Then, only utterances that represented the speakers "normal" speaking style was selected. For example, utterances where the speaker put on different voices to portray a third person, such as his wife or friends, were excluded. Utterances with word fragments, mispronunciations, heavily reduced pronunciations, mumbling and laughter were also excluded, to allow standard forced alignment (Young et al., 2006). However, we emphasise that the remaining utterances were still rich in conversational speech phenomena, in particular backchannels, discourse markers and filled pauses. An example is shown below:

> *"yeah it's it's a significant amount of swelling um more than like I'd say a bruise"*

4.3. Forced Alignment and Analysis

The forced alignment and linguistic "front-end" analysis of the transcribed conversational speech were made with the CereVoice (Aylett and Pidcock, 2007) speech synthesis system. The segmentation procedure follows the forced alignment method outlined in Young et al. (2006). In addition to detecting and aligning utterance internal pauses, the system made use of a rich system of pronunciation variants, with full and reduced pronunciation variants for many frequent English words[2], e.g. *and* can be pronounced fully (ænd) or reduced (ən), *but* can be pronounced fully (bʌt) or reduced (bət), and *the* can be pronounced fully (θiː) or reduced (θə). Many other words also had pronunciation variants, e.g. in *because* and *'cause*, where the vowel quality was decided automatically based on the specific speaker's usage.

Although these pronunciation variants gave a better match to pronunciations in the conversational speech, forced alignment initiated with just the conversational speech did not provide a sufficiently accurate alignment. By utilising phoneme models trained from additional read

[2]All phonemes in this paper are denoted with IPA symbols.

Table 1: A comparative analysis of phone alignment accuracy between read aloud and conversational speech. Ten randomly selected utterances with a total of approximately 300 phones were used.

	# phones	total error	max error/phone
Conversational	347	2085 ms	800 ms
Read aloud	313	710 ms	80 ms

aloud speech recorded for phonetic coverage from the same speaker, the forced alignment of the conversational speech was improved (Andersson et al., 2010a). This alignment and linguistic analysis were also used for the HMM-based voices in this paper.

To evaluate the alignment of the conversational speech, ten randomly selected utterances with a total of approximately 300 phones for the conversational and read aloud speech were compared with manually annotated results. Results are shown in Table 1. This shows that total alignment error in the conversational speech is about three times that of the read aloud speech. However this error is mostly concentrated in a few specific segments. 1500ms of the total 2085ms error can be accounted for by only two segments from one particular sentence. In general, excluding gross alignment errors, the alignment of the conversational speech was found to be as accurate as the read aloud speech.

4.4. Context Dependent Phonemes

The context dependent phonemes define the segmental and prosodic categories and dependencies in speech, for both the training and generation parts of HMM-based speech synthesis, and were generated with the CereVoice system from the text analysis and corresponding forced aligned utterances. CereVoice's contexts were based on the contexts in Tokuda et al. (2002) 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)

Although the contexts did not include explicit representation of the backchannels, discourse markers or filled pauses, the context specifications implicitly identified many important characteristics. The quinphone context encapsulated many of the discourse markers and filled pauses, e.g. *yeah*, *you know* or *oh yeah*, together with their frequent utterance initial or final positions. The quinphone context was also large enough to cover a filled pause 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 hesitation and discourse function. The contexts with counts and phrase positions should also be able to capture segmental and prosodic differences between e.g. *yeah* as a stand alone backchannel, the confirmation *yeah yeah yeah*, or in the longer utterance *yeah I feel kind of dirty afterwards*.

The current contexts did not, by any means, capture all the important characteristics, and better and more accurate text/speech analysis and representation of conversational speech phenomena is an important problem for HMM-based speech synthesis. However, the current contexts allowed us to establish a baseline for the continued research of conversational speech in HMM-based speech synthesis and to make a direct comparison where the only varying factor is the underlying data.

Table 2: Overview of the conversational and read aloud data. The duration shows the amount of phonetic material, including or excluding utterance internal silent pauses. The quinphone types include silences, but not lexical stress.

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

5. Comparing Conversational and Read Aloud Data

5.1. Word distribution

In addition to the conversational data, we used neutrally read aloud sentences available from the same male American speaker, recorded in the same studio, with the same microphone, around the same time as the conversation (Andersson et al., 2010a). The read aloud sentences came from a wide range of text genres, including news, weather reports, and "conversation". An overview of the conversational and read aloud data is shown in Table 2. In terms of duration of usable utterances there was more read aloud data than conversational data, however the distribution of useful characteristics in each set were quite different. Table 3 shows the twenty most frequent words in the conversational and read aloud data. It is no surprise that short function words, such as *the*, *a*, *of* or *to*, were

Table 3: The 20 most frequent words in the conversational and read aloud data. Non-overlapping words between the two columns are bold faced.

	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

frequent in both the read aloud and conversational data. More interestingly, 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 in 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 the utterances in the conversational speech data were backchannels consisting of a single word (e.g. 339 *yeah*, 167 *right* and 54 *okay*), but the discourse markers and filled pauses are mainly integrated with propositional content in longer utterances, and as Table 4 shows, often occur in the vicinity of the phrase or utterance boundaries, hence represent our speaker's means of starting, ending or keeping a turn.

5.2. Prosodic Properties

HMM-based speech synthesis generally assume recordings of consistently spoken material, where the main difference between utterances is the sequences of phonemes. Conversational speech however, has more segmental and prosodic variation. An example of the consistency in carefully read aloud speech and the variation in conversational speech is the speaking rate, showed in Figure 1.

Whereas some of the variation in conversational speech can probably be attributed to less careful delivery, much of the variation has other explanations related to the language used (Shriberg and Stolcke, 1996; Bell et al., 2003; Aylett and Turk, 2006). Figure 2 shows that whereas our speaker had approximately the same pitch range and F0 distribution when reading aloud and when speaking in a conversation, the conversation had more variation in utterance final F0, probably because of more variation in speech acts (questions, confirmations, etc.) and speaker state (enthusiastic, doubtful, polite, etc.).

Table 4: The 30 most frequent trigrams in the conversational data. Excluding one word backchannels, but including utterance beginning/end as "sil", and utterance internal short pauses as "sp". In total 82 trigrams in the conversational data occurred 10 times or more. In comparison, only seven trigrams in the read aloud data occurred more than 10 times.

frequency	trigram	frequency	trigram
124	sil_yeah_sp	27	sp_so_sil
118	sp_you_know	23	sp_uh_sp
68	sil_um_sp	23	sp_yeah_sp
68	sil_you_know	20	you_know_and
53	yeah_sp_yeah	20	sil_and_then
46	you_know_sp	19	uh_sp_yeah
43	you_know_what	19	sil_and_I
38	know_what_I	18	I_mean_sil
38	you_know_sil	18	yeah_yeah_sil
37	a_lot_of	17	you_know_I
37	sp_um_sp	17	but_uh_sp
37	sp_yeah_sil	16	sp_and_uh
36	sp_um_sil	16	and_uh_sp
36	what_I_mean	16	sp_and_then
27	sil_yeah_I	16	sp_yeah_yeah

5.3. Segmental Properties

Figure 3 shows the average centre frequencies of automatically extracted first and second formants of vowels in the conversational and read aloud data. In this figure, we cannot see a clear tendency to a reduced vowel space in the conversational speech, contrary to Nakamura et al. (2008). There were two main reasons for this. The first reason is that the data did not contain many heavily reduced pronunciations. The second reason is that the forced alignment process have explicitly dealt with vowel reduction and have assigned a schwa in reduced variants of *the*, *you*, *but*, etc. However, there were still some reduction tendencies observable, and an example of different vowel formant values in fully pronounced and reduced *it* is shown in Figure 4.

We can also see some differences related to language differences or lexicon errors. The vowel in the filled pauses was stipulated in the pronunciation lexicon as an /ʌ/, but as Figure 3 shows, this was not correct, and was the main reason for the difference between read aloud and conversational /ʌ/. After we manually verified the automatically extracted formant values for the vowel /u/, we found that the formants were not well estimated for the word *you*, and probably not for other short function words, like *to* or *do*, which had a higher proportion in the conversational data.

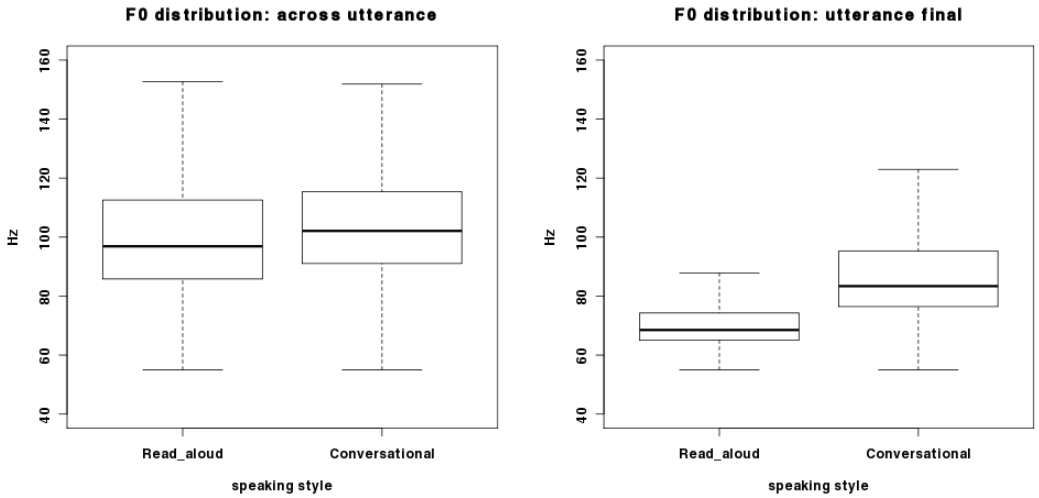

Figure 2: F0 distribution in read aloud and conversational speech. Left figure shows F0 variation across the whole utterance. Right figure shows F0 variation at the end of utterances. Due to uncertainties of F0 at the end of utterances, the utterance final F0 was measured at the 10th last voiced frame, frame length was 5ms.

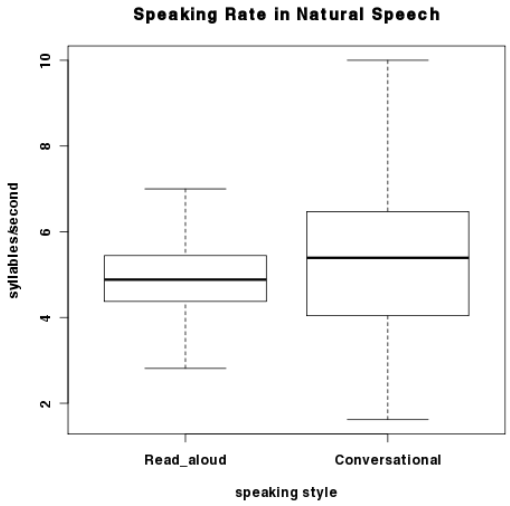

Figure 1: Speaking rate for utterances with 5-10 words in the conversational and read aloud data. The solid line is the median, box borders show the upper and lower quartiles, and the whiskers are drawn to 1.5 times the inter-quartile range. The speaking rate of the conversational and read aloud data 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, and it is questionable if speaking rate is a relevant measure for backchannels, therefore the speech rate was only measured for utterances that were five to ten words long.

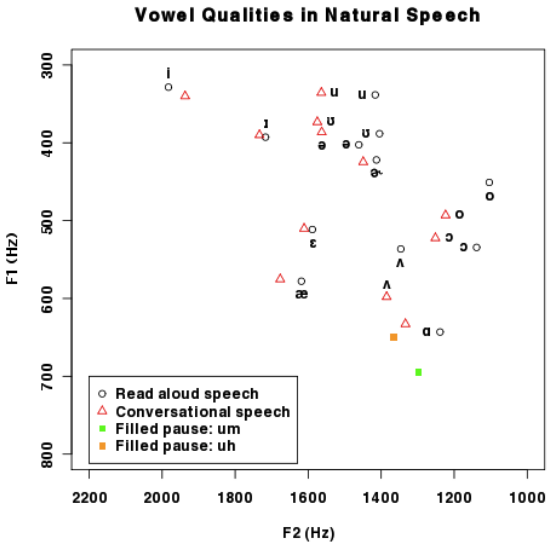

Figure 3: Mean formant values (F1 and F2) for American English monophthongs, denoted with IPA symbols, in read aloud and conversational speech. The mean formant values for the two filled pause types (*um* and *uh*) in the conversational speech are also plotted.

6. Conversational and Read Aloud Synthetic Voices

The context dependent phonemes introduced in Section 4.4 for the read aloud and conversational speech, were used to build one "spontaneous" and one "read aloud" synthetic voice with the HTS system described in Section 3. These voices are henceforth referred to as *Spontaneous HTS* and *Read Aloud HTS*, respectively.

The size of the clustered decision trees reflects the amount and complexity of the speech data. Table 5 shows that despite less data for the Spontaneous HTS than the Read Aloud HTS voice the clustered duration tree was larger for the Spontaneous HTS due to more variation needing to be accounted for. Unlike, for example, the mel-cepstral tree where the Read Aloud HTS tree was larger due to more data and better quinphone coverage.

6.1. Blending Read Aloud and Conversational Speech

Our first impression of the quality of the Spontaneous HTS 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,

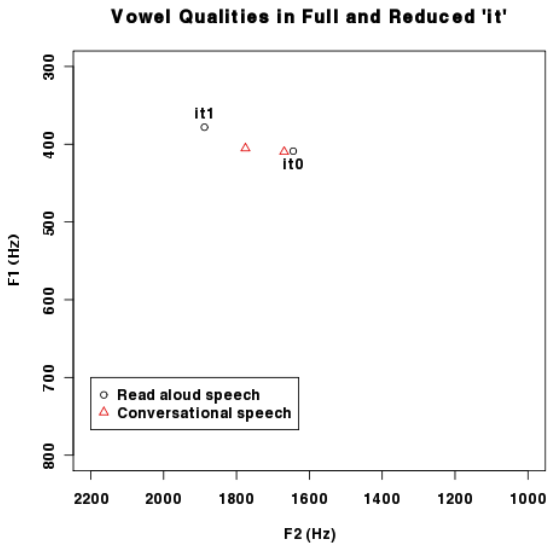

Figure 4: Mean formant values (F1 and F2) for fully pronounced and reduced vowel in the word *it*. Represented as it1 (full) and it0 (reduced) in the figure.

Table 5: Number of leaf nodes in the clustered duration, logF0, mel-cepstral and aperiodicity trees, for the Spontaneous HTS (SP) and Read Aloud HTS (RD) voices. The ratio(SP/RD) shows the relative tree sizes.

	Spon. (SP)	Read (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

the conversational and read aloud data were blended in the training and clustering of HMM-based models with a method previously used to blend and preserve different "emotional" speaking styles (Yamagishi et al., 2005).

All the conversational and read aloud data were pooled in training, and an additional context: speaking style (spontaneous or read), was added to the context dependent phoneme descriptions in Section 4.4. In the training of the context dependent HMM-based models, the speaking style context was then available as a question in the decision tree based clustering, and was automatically selected in the clustering to share mutual and sparse phonetic properties between conversational and read aloud speech, while avoiding to share frequent and distinguishing characteristics. The speaking style context was automatically selected as an important feature throughout the clustering process. For example, in the clustering of the duration tree, a split was made almost immediately based on the difference in duration of the syllable nucleus in conversational and read aloud speech, whereas for the excitation and spectral part the sharing or splitting seemed to be more complex.

During synthesis with this voice one of the speaking styles was selected by setting the speaking style context to either spontaneous or read aloud, and then speech parameters were generated. Henceforth, utterances generated in this way are referred to as from the *Blend.Spon* voice and *Blend.Read* voice, respectively.

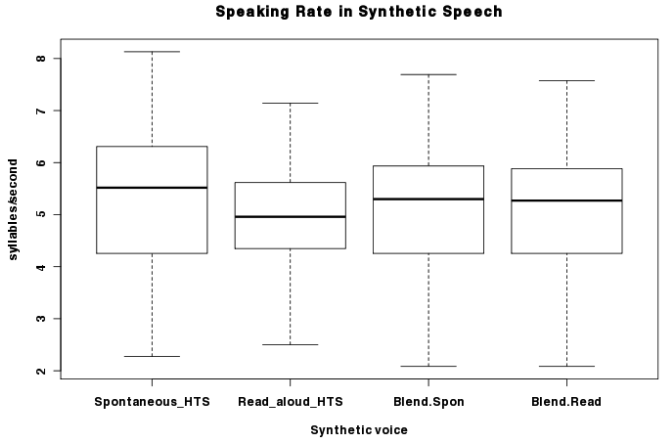

Figure 6: Speaking rate for 169 synthetic utterances with the same phonemic sequence, synthesised with four different synthetic voices.

6.2. Test Set for the Synthetic Voices

A test set of synthetic speech was generated from each of the synthetic voices: the Spontaneous HTS, the Read Aloud HTS, the Blend.Spon and the Blend.Read voice. The context dependent phonemes for the synthetic speech in the test set were obtained from unused transcripts of the same conversation and speaker as in Section 4. 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 speaking and it was rich in conversational speech phenomena; 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 were rich in conversational phenomena, and had identical phonemic sequences and linguistic analysis, hence, allowing a linguistically balanced acoustic comparison.

6.3. Segmental and Prosodic Properties

In Section 5.2 and 5.3 we showed some segmental and prosodic differences between read aloud and conversational speech. In this section we will show segmental and prosodic differences between the synthetic voices built with either conversational or read aloud speech, and the blended voice built with both.

Figure 5 shows a comparison of the first two formants in the synthetic speech. For the Spontaneous HTS and the Read Aloud HTS 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 to a reduced vowel space in the Spontaneous HTS compared to the Read Aloud HTS, and again the manually removed heavily reduced pronunciations from the conversational data probably contributed to this. We can also

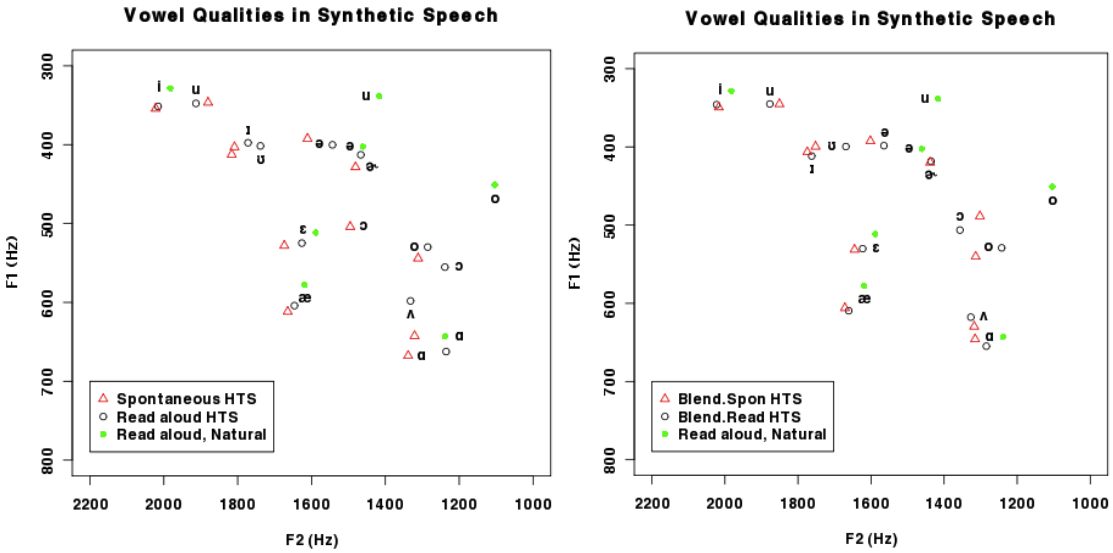

Figure 5: Mean formant values (F1 and F2) for American English monophthongs, denoted with IPA symbols, in 169 utterances with the same phonemic sequence, synthesised with four different synthetic voices. Some of the natural vowels from Figure 3 are provided as a reference.

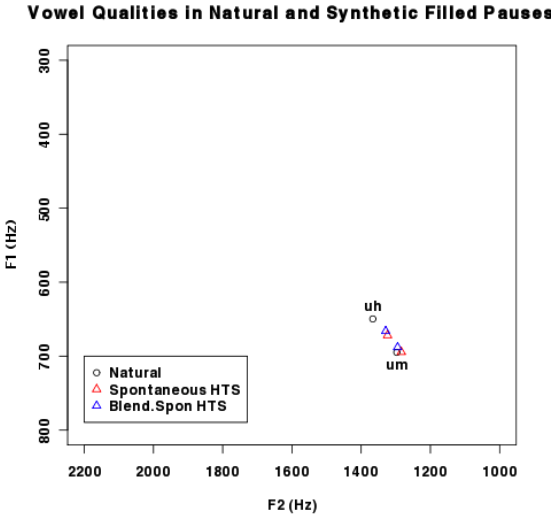

Figure 8: Vowel quality of filled pauses in conversational speech, Spontaneous HTS and Blend.Spon. (Vowel qualities of the filled pauses in Read Aloud HTS are not plotted, but were more different; um F1:661/F2:1342, uh F1:589/F2:1399.)

see that the extracted vowel formants were closer to each other for the blended voice than in the Spontaneous HTS and Read Aloud HTS or in the natural read aloud and conversational speech[3]. This pattern was also preserved in the synthetic speaking rate, shown in Figure 6, where the Spontaneous HTS and Read Aloud HTS 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 Spontaneous HTS, as well as the Blend.Spon voice, and different from the vowel quality and duration in the Read Aloud HTS (see Figure 7 and Figure 8). The duration of *uh* in the Read Aloud HTS had more similarity than the *um* to the natural filled pauses, but there were no filled pauses in the read aloud speech data, and 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.".

In general, there was more variation in the natural speech than in either of the synthetic voices. But, Figure 10 shows an utterance initial filled pause where the Spontaneous HTS 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 Read Aloud HTS were different and did not convey any hesitation and therefore did not sound like a filled pause. Similarly, many discourse markers were also generally well preserved in both the Spontaneous HTS and Blend.Spon. Figure 9 shows an utterance initial *yeah*, followed by a short pause, from natural and synthetic speech, where the Spontaneous HTS had segmental and prosodic properties that were similar to the natural reference sample, whereas the *yeah* from the Read Aloud HTS had completely different shape of the F0 contour, longer duration of the vowel part of the *yeah*, and despite that the phonemic sequence was intelligible, it came across as almost meaningless.

7. Perceptual Evaluation

7.1. Listening Test Design

People can distinguish perceptually between natural spontaneous and natural read aloud speech (Blaauw, 1994;

[3]We also see that the /u/ vowel in non-reduced *you*, *to*, *do* and *doing* where due to co-articulation F2 starts off high, were difficult to automatically extract formants from in all the synthetic voices. In the /ɔ/ vowel, as in *more*, *long* or *talk*, the closeness of F1 and F2 made the automatic formant extraction unreliable.

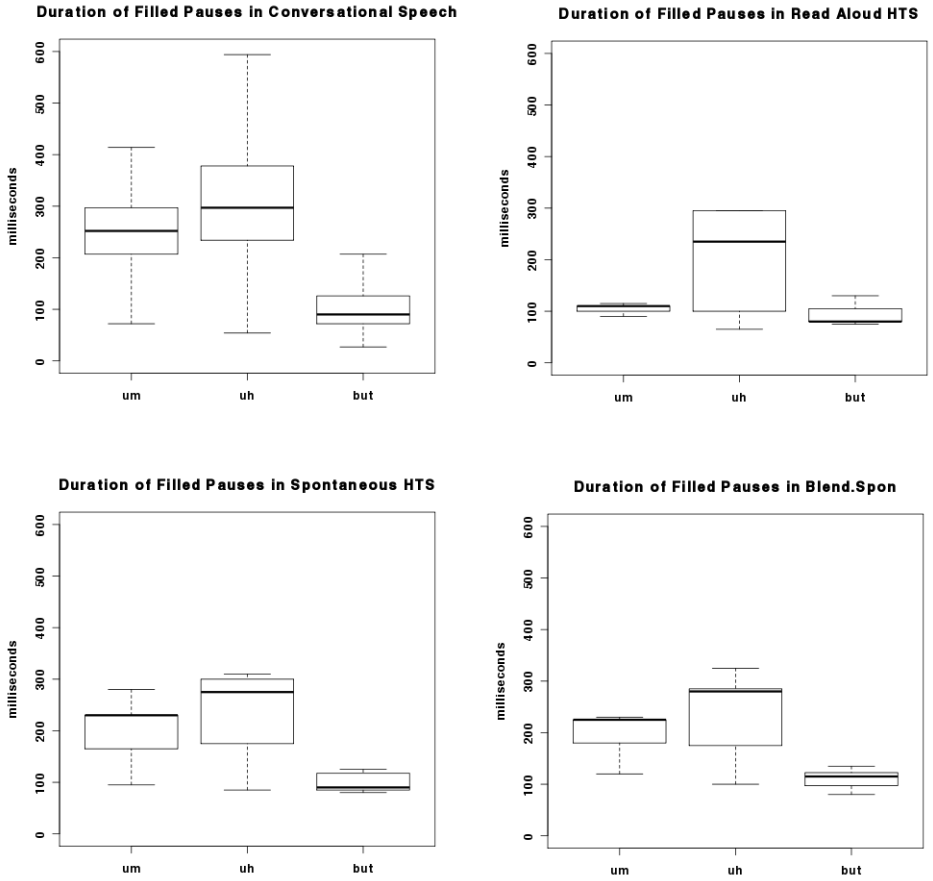

Figure 7: Duration of the vowel in the filled pauses (*um* and *uh*), and in the reference word *but*, for natural and synthetic speech. *But* was used as reference because it was represented in the lexicon as having the same vowel quality as the filled pauses, and existed in both the natural and synthetic speech.

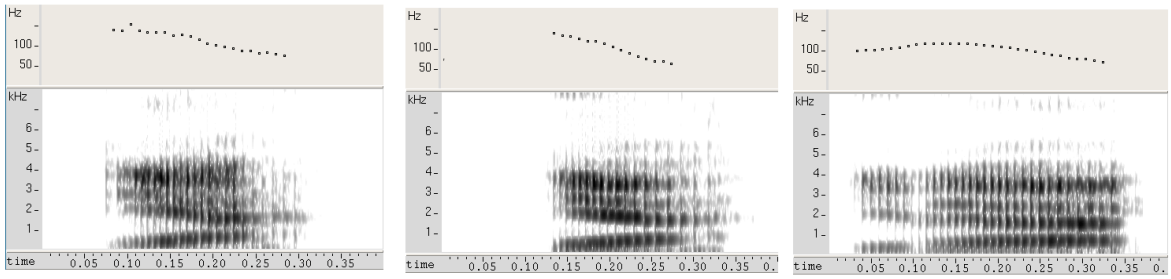

Figure 9: A *yeah* in the "same" utterance: natural (left), Spontaneous HTS (mid), and Read Aloud HTS (right).

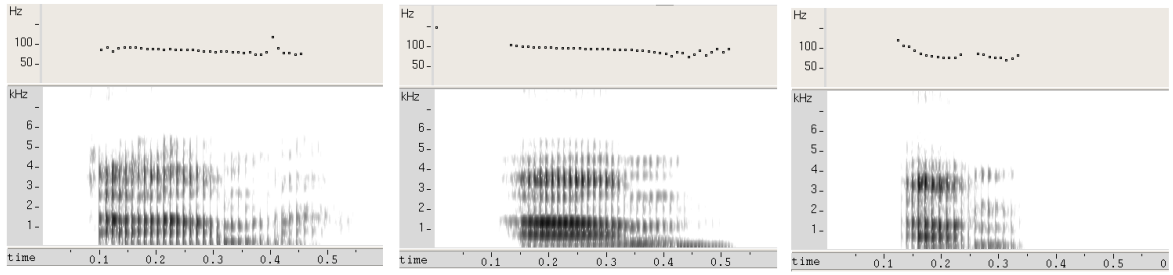

Figure 10: The filled pause *um* in the "same" utterance: natural (left), Spontaneous HTS (mid), and Read Aloud HTS (right).

Laan, 1997) and we designed a listening test to evaluate whether such a distinction could also be observed for HMM-based synthetic voices built with spontaneous conversational or read aloud speech. In particular, we designed the evaluation to investigate the impact on speech quality and speaking style on the integration of conversational characteristics: discourse markers and filled pauses, with propositional content.

"Naturalness" is conventionally used in speech synthesis to evaluate speech quality, but evaluating a spontaneous or conversational speaking style has been less explored. We suspect that when listeners are asked to judge the quality of synthetic speech using a positive vs negative distinction they do so in a quite general way, not paying particular attention to the specific feature they have been asked to judge. To investigate this issue further, the listening test was designed to determine whether naturalness could be evaluated separately from speaking style. To do so, listeners were divided into two groups and the two groups were given one criteria each to evaluate, quality or speaking style.

One group of listeners was requested to evaluate: "Which utterance sounds more like natural speech?". Another group of listeners was requested to evaluate "Which utterance has a more conversational speaking style?". The listeners 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.".

Test sentences for the listening test were randomly selected from the set of the 169 synthetic utterances, 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. All test sentences are shown in Table 6.

To evaluate the integration of discourse markers and filled pauses with propositional content in synthetic speech the listening test compared pairs of utterances synthesised with the Spontaneous HTS to utterances synthesised with the Read Aloud HTS. To evaluate the contribution of discourse markers and filled pauses on quality and speaking style, utterances with these conversational characteristics synthesised with the Blend.Spon were compared to more conventional text-to-speech utterances where discourse markers, filled pauses and disfluencies were removed from the original sentence and synthesised with the Blend.Read voice.

To avoid a scenario where it was obvious from text alone that the discourse markers and filled pauses had been removed from one of the utterances, we always compared utterances with completely different lexical content. For example: if we had compared A) *so let's see, but um, yeah, nothing exciting*, to B) *let's see, but nothing exciting*, listeners could quite easily identify that one utterance had the same 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 lexical differences would make it harder to identify that we just removed a few words, and hence evaluate speaking style and not text style.

7.2. Listening Test Results

The listening results in this section were reported in Andersson et al. (2010b) and summarised here. The result of the first listening test (to evaluate the integration of discourse markers and filled pauses) is shown in Figure 11. We can see that the perceptual judgements were significantly in favour of the Spontaneous HTS, due to the better realisations of the discourse markers and filled pauses, and thereby also a better utterance prosody. The result of the second listening test (to compare utterances with and without conversational characteristics) is shown in Figure 12. We see that when these conversational characteristics were removed from the test sentences synthesised with the Blend.Read voice, and compared to sentences with discourse markers, filled pauses and disfluencies synthesised with the Blend.Spon voice, the results were more in favour of the Blend.Read voice. This is an interesting tendency and therefore, in the following sections, we will take a closer look at why removing a few words had such an impact on the perception of naturalness and speaking style.

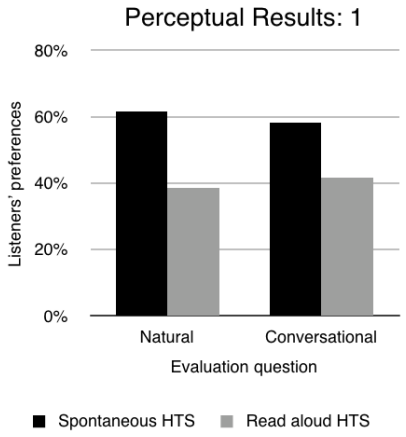

Figure 11: The bars show the percentages of the listeners' preferences for naturalness and conversational style when comparing the Spontaneous HTS to the Read Aloud HTS when synthesising utterances with discourse markers and filled pauses.

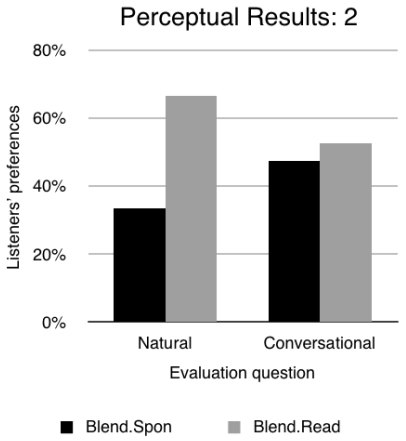

Figure 12: The bars show the percentages of the listeners' preferences for naturalness and conversational style when comparing utterances with conversational characteristics (Blend.Spon) to more fluent utterances (Blend.Read).

Table 6: All the utterance pairs in the perceptual evaluation. The pairs for the Blend.Spon and Blend.Read evaluation are shown in the Text boxes. The utterances for the Spontaneous HTS are the same as for the Blend.Spon. The utterances for the Read Aloud HTS can be derived by replacing the Blend.Read utterance, with the next Blend.Spon utterance, e.g. the Read Aloud HTS utterance for pair 3 would be *yeah, x-men is cool, yeah* instead of *x-men is cool*. Commas indicate where utterance internal silences were located.

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

7.3. Naturalness

Figure 13 and Figure 14 show the listeners' perceived naturalness for individual utterances behind the results summarised in Figure 11 and Figure 12. By listening to the utterances we identified a few factors that probably contributed to the perceived difference in naturalness between utterances in Figure 13 and Figure 14.

Some factors were easily identified in Figure 14: e.g. the prominent local pitch movement errors in the Blend.Spon utterances 10 and 11, the awkward prosody in utterance 15, or the too prominent word repetition in utterance 5. This was probably to some extent a reflection of our underspecified analysis and representation of segmental and prosodic properties in conversational speech, including segmentation, prosodic phrasing and disfluencies. But, the prominent word repetition in utterance 5 was prominent also in the original natural utterance, and the listeners judgements were perhaps influenced by the presence of an audible disfluency, a factor that we discuss also in the next paragraph.

An important general tendency in the perceived naturalness between Figure 13 and 14 was that 1) the conversational characteristics sounded bad with the Read Aloud HTS and 2) that when the discourse markers, filled pauses and disfluencies were removed in the sentences synthesised, it made the Blend.Read utterances sound substantially better. The removal of discourse markers, filled pauses and disfluencies made many utterances more grammatical and more fluent than the conversational utterances, e.g. utterance pairs 2, 3, 4, 9, and 14, which could have contributed to making the perceived differences in naturalness larger than they were and contribut to the differences for these utterance pairs between Figure 13 and Figure 14.

7.4. Conversational Speaking Style

The perceptual evaluation was designed to see if we could evaluate speaking style separately from naturalness. The questions about naturalness and speaking style were therefore asked to separate groups of listeners.

The speaking style results in Figure 11 was significantly in favour of the Spontaneous HTS, but so was the naturalness, and the correlation between them was significant ($\rho = 0.72,\ p << 0.05$). Our interpretation was that the difference between the voices in Figure 11 was mainly a re-

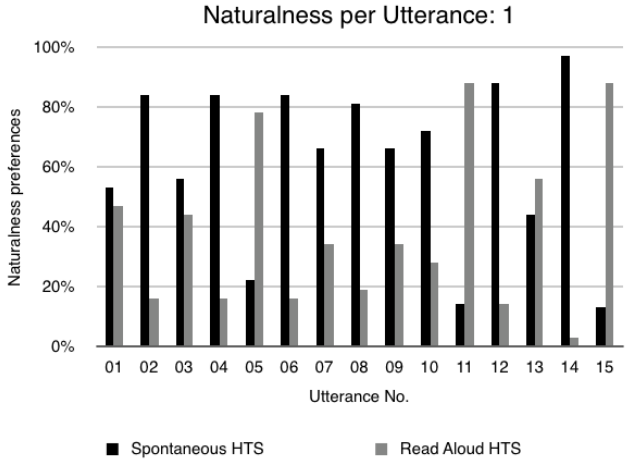

Figure 13: Listeners' perception of naturalness for individual utterances when comparing sentences with discourse markers and filled pauses synthesised with the Spontaneous HTS or Read Aloud HTS voices.

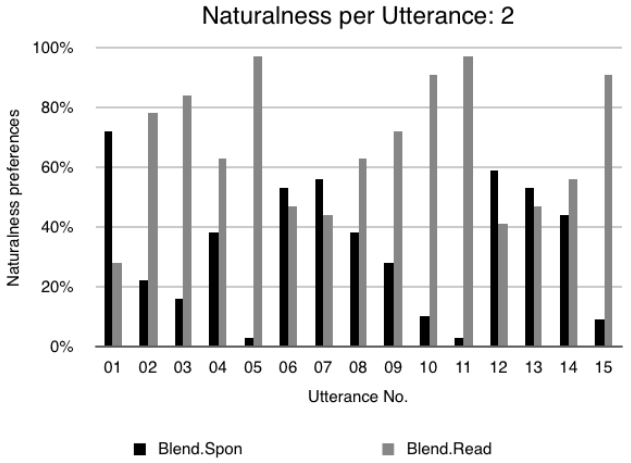

Figure 14: Listeners' perception of naturalness for individual utterances when comparing sentences synthesised with the blended voice. The Blend.Spon bar shows preference for utterances with conversational characteristics, and the Blend.Read bar shows preference for more fluent utterances.

flection of the difference in naturalness rather than speaking style.

In Figure 12 the charts for naturalness and speaking style were different, but there was no significant difference between the perceived speaking style for the two voices. However, correlation between the two groups' perception of naturalness and speaking style was stronger ($\rho = 0.86,\ p << 0.05$), as shown in Figure 15. 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 probably contributed to making the evaluation of speaking style more difficult for the listeners.

Figure 16 shows individual listeners' perception of conversational speaking style, and we can see that there were

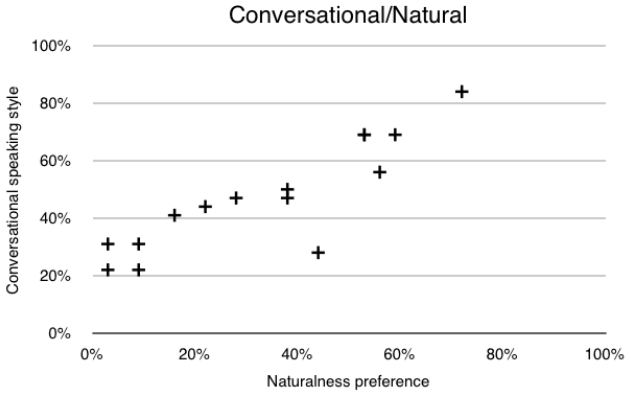

Figure 15: Plot of the listeners' preferences of naturalness and conversational speaking style for the Blend.Spon voice. Spearman's rho showed significant ($p << 0.05$) correlation of 0.86

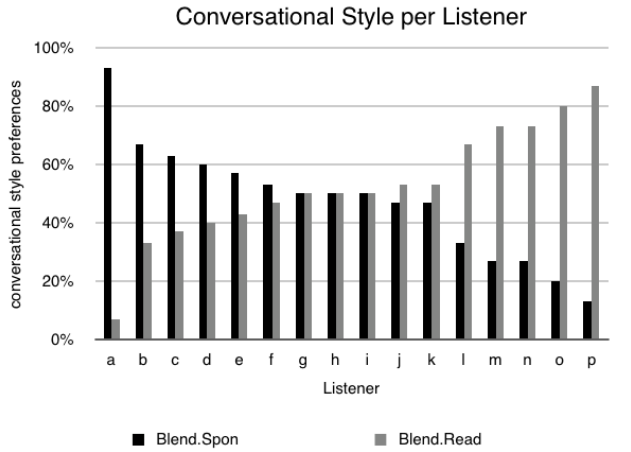

Figure 16: Individual listeners' perception of conversational speaking style when comparing sentences synthesised with the blended voice. The Blend.Spon bar shows preference for utterances with discourse markers, filled pauses and disfluencies, and the Blend.Read bar shows preference for more fluent utterances.

at least two different interpretations of speaking style, where listeners *a-d* have interpreted speaking style differently than listeners *l-p*.

8. Conclusions

We have shown that speech from a spontaneous conversation was instrumental for building an HMM-based synthetic voice that could integrate discourse markers, filled pauses and propositional content into more natural sounding utterances than an HMM-based voice built with carefully read aloud sentences. It was able to capture distinguishing segmental and prosodic properties of conversational speech phenomena, despite worse phoneme alignment and underspecified and erroneous linguistic analysis.

The strong correlation in the perceptual evaluation between naturalness and speaking style showed that it is difficult to separately evaluate these aspects in synthetic speech. A better alternative for future evaluations of conversational speech synthesis than naturalness or speaking style, could be to evaluate agreement, disagreement, hes-

itation, etc. in a discourse context, as in Lasarcyk and Wollerman (2010).

The quality and expressiveness of the synthetic voices achieved from a moderate amount of conversational speech data provide an encouraging start for the continued research of conversational speech in HMM-based speech synthesis. More sophisticated analysis and representation of conversational speech phenomena, as well as developments of the training/generation framework of HMM-based speech synthesis, would probably allow further improvements to synthesising the rich variation of speech phenomena in human conversation.

Acknowledgements

The authors are grateful to David Traum and Kallirroi Georgila at the USC Institute for Creative Technologies (http://ict.usc.edu) for making the speech data available to us. The first author is supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

References

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

J. Adell, A. Bonafonte, and D Escudero-Mancebo. Modelling filled pauses prosody to synthesise disfluent speech. In *Speech Prosody*, Chicago, U.S.A., 2010.

S. Andersson, K. Georgila, D. Traum, M.P. Aylett, and R.A.J. Clark. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Speech Prosody*, Chicago, USA, 2010a.

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

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

M.P. Aylett and C.J. Pidcock. The CereVoice characterful speech synthesiser SDK. In *AISB'07*, Newcastle Upon Tyne, U.K., April 2007.

L. Badino, S. Andersson, J. Yamagishi, and R.A.J. Clark. Identification of contrast and its emphatic realisation in HMM based speech synthesis. In *Proc. Interspeech*, Brighton, UK, 2009.

A. Bell, D. Jurafsky, E. Fossler-Lussier, C. Girand, M. Gregory, and D. Gildea. 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, 2003.

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

D. Cadic and L. Segalen. Paralinguistic elements in speech synthesis. In *Interspeech*, Brisbane, Australia, 2008.

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

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

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

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

K. Gustafsson and K. Sjölander. Voice creation for conversational fairy-tale characters. In *SSW5-2004*, pages 145–150, Pittsburgh, U.S.A., 2004.

D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. Lexical, prosodic, and syntactic cues for dialog acts. In *ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, 1998.

H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné. 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*, 27:187–207, 1999.

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

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

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

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

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

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

J. Odell. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, 1995. PhD Thesis.

D. O'Shaughnessy. Recognition of hesitations in spontaneous speech. In *ICASSP*, pages 521–524, San Fransisco, USA, 1992.

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

D. Schiffrin. *Discourse Markers*. Cambridge University Press, 1987.

E. Shriberg. Phonetic consequences of speech disfluency. In *Proc. of International Congress of Phonetic Science*, pages 619–622, San Fransisco, U.S.A., 1999.

E. Shriberg and A. Stolcke. Word predictability after hesitations: A corpus-based study. In *Proc. of ICSLP*, Philadelphia, U.S.A., 1996.

S. Sundaram and S. Narayanan. Spoken language synthesis: Experiments in synthesis of spontaneous dialogues. In *Proc. of 2002 IEEE SSW*, Santa Monica, USA, 2002.

K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to English. In *Proc. of 2002 IEEE SSW*, Santa Monica, U.S.A., 2002.

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

W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer-Verlag Berlin Heidelberg, 2000.

J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. Acoustic modelling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Trans. Information and Systems*, E88-D(No. 3), 2005.

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

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

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