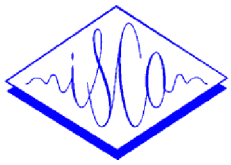
ISCA Archive
http://www.isca-speech.org/archive



7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Utilising Spontaneous Conversational Speech in HMM-Based Speech Synthesis

Sebastian Andersson, Junichi Yamagishi, Robert Clark

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
J.S.Andersson@sms.ed.ac.uk

Abstract

Spontaneous conversational speech has many characteristics that are currently not well modelled in unit selection and HMM-based speech synthesis. But in order to build synthetic voices more suitable for interaction we need data that exhibits more conversational characteristics than the generally used read aloud sentences. In this paper we will show how carefully selected utterances from a spontaneous conversation was instrumental for building an HMM-based synthetic voices with more natural sounding conversational characteristics than a voice based on carefully read aloud sentences. We also investigated a style blending technique as a solution to the inherent problem of phonetic coverage in spontaneous speech data. But the lack of an appropriate representation of spontaneous speech phenomena probably contributed to results showing that we could not yet compete with the speech quality achieved for grammatical sentences.

Index Terms: HMM, speech synthesis, spontaneous, conversation, lexical fillers, filled pauses

1. Introduction

Unit selection and HMM-based speech synthesis has achieved high levels of naturalness and intelligibility when synthesising read aloud sentences [1]. For many applications, an intelligible read aloud speaking style is sufficient to provide a user with relevant information. But speaking is different from reading aloud and applications that require conversational interaction, e.g. believable virtual characters [2] or speech-to-speech translation [3], need synthetic voices that can synthesise turn-taking behaviour, provide back-channels and express agreement, disagreement, hesitation, et cetera.

Unit selection and HMM-based synthetic voices are generally built from recordings of several thousands of carefully read aloud sentences that enables synthesis of sentences that are not pre-recorded. But the recorded sentences also determines the speaking style of the voice and gives it a read aloud character. In HMM-based speech synthesis the recordings of emotional speaking styles was instrumental for synthesising emotional speaking styles [4], the recordings of emphatic accents was instrumental for synthesising emphatic accents [5], and in order to build synthetic voices that are more suitable for conversation we need to build voices from data that exhibits more conversational characteristics.

In this paper we report on work with eliciting and selecting speech from a spontaneous conversation and show that this data was instrumental for synthesising distinguishing characteristics between spontaneous and read aloud speech in HMM-based speech synthesis.

1.1. Filled Pauses, Lexical Fillers and Conversational "Grunts"

Spontaneous speech has many properties generally absent from carefully read aloud speech, such as word fragments or heavily reduced pronunciation. In this work we have largely avoided these by the selection of spontaneous speech material described in section 2.1, to allow us to focus on the speech phenomena that are important for regulating turn-taking and express affective content.

An inclusive description of these phenomena as "wrappers" around propositional content was given in [6] and an example from our data is given below with wrappers in italics and 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** "

The wrapper category can be further divided into filled pauses, lexical fillers and conversational "grunts" based on their phonetic properties. In conversation these fillers and grunts provide efficient means to regulate the flow of the conversation (turn-taking, back-channels) and express affective content (e.g. agreement, disagreement and hesitation) [6, 7].

In an analysis of the phonetic properties of conversational "grunts" in American English [7] identified a small set of frequently occurring phonetic features, e.g. /h/, /m/, schwa, and creaky or breathy voice qualities. The non-lexical nature of the grunts makes them difficult to exemplify in text, but a few examples with fairly recognisable meaning are *"uh-huh"* and *"hmm"*.

The lexical fillers, as the name suggest, consists of more lexicalised items, but their phonetic properties determines their function as fillers and [7] argues that some (e.g. *"yeah"*, *"right"*) are more grunt-like than others (e.g. *"you know"*). A detailed analysis of stand-alone *"so"* showed how the phonetic differences was related to different turn-taking functions [8].

The filled pauses (*"um"* and *"uh"*) are generally regarded as a hesitation phenomena. The vowel quality is often close to a schwa, but can also have other vowel qualities. Their duration is on average much longer than vowels elsewhere in an utterance, and as a hesitation phenomena they are associated with a prolongation of at least the preceding syllable [9].

1.2. Related Work

Most work on utilising spontaneous or conversational speech resources for synthesis have been done with other synthesis techniques than HMM-based speech synthesis, e.g. unit selection [10, 11], limited domain synthesis [12] or phrase level selection from a very large corpus [13]. But whereas [14] did use spontaneous speech in HMM-based speech synthesis to model pronunciation variation, [10, 11, 12, 13] focused on the "fillers" or "grunts" that are the focus also of this paper. The work in

[10, 12, 13] all used concatenative methods where the phonetic detail of the fillers and grunts from spontaneous speech was preserved, but there were no means of generating fillers and grunts with unseen or combined meanings. HMM-based speech synthesis offers a framework that potentially would allow training and generation of unseen fillers and grunts, and although this would require solutions beyond the current synthesis framework an essential requirement is that there is appropriate data available and in this paper we will investigate the consequences of utilising data from a spontaneous conversation in current HMM-based speech synthesis to begin identifying future requirements.

2. Spontaneous Conversational Speech Data

The spontaneous conversational speech data in this paper was recorded for use in speech synthesis and was previously used in [10]. This section gives a more detailed description of that data and contrast it with read aloud sentences recorded for phonetic coverage with the same voice talent, microphone, and in the same studio.

Approximately seven hours of unconstrained conversation was recorded between the voice talent and the first author over a period of three days. The voice talent was an American male from Texas in his late thirties, and the author was a male non-native English speaker in his late twenties. The voice talent was positioned inside a recording booth and the author was positioned outside it. They communicated via headphones and microphones, but had eye-contact through a window. The speech from the voice talent and author were recorded on separate tracks, but technical problems resulted in the speech of the author only being captured with very low gain.

The conversation was unconstrained, but mainly focused around the voice talent's work as an actor, former sports career and life in general in the U.S. He was aware of the overall goal of the conversation, i.e. to use his speech for synthesis. Although the intent was to elicit spontaneous conversational speech in a natural discourse, the voice talent was given some feedback on his speech behaviour and was requested to not put on so many different "voices" when portraying a third person, something he frequently forgot.

2.1. Transcription and Selection

The speech of the voice talent was transcribed orthographically and aligned at the utterance level. The motivation for an orthographic transcription was that it left the precise meaning of fillers, grunts, back-channels, pitch contours et cetera, underspecified while still identifying a token level suitable for subsequent manual or automatic processing.

The transcription and selection of speech aimed to obtain data that represented the speakers "normal" speaking style and specific language use. To get data that in some sense could be considered his consistent spontaneous conversational speaking style. Utterances where the speaker put on different voices to portray a third person, such as his wife or friends, were therefore excluded.

Para-linguistics, and in particular perhaps laughter, is part of conversational interaction, but were excluded to limit the range of phenomena needed to deal with. We wanted data that allowed us to focus on the communicative units specific to conversational speech, and although the relation between words and subword units is an important problem we excluded utterances with word fragments, mispronunciations and heavily reduced pronunciations.

In total we obtained 2120 utterances, approximately 75min of phonetic material (without silent pauses), that were rich in spontaneous speech phenomena, in particular back-channels, filled pauses and lexical fillers, but free from word fragments, mumbling, heavily reduced pronunciations and para-linguistics. An example is shown below:

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

Utterance internal silent pauses was later detected through forced alignment (see sec. 2.4).

2.2. Spontaneous versus Read Aloud Data

Several studies have showed that listeners can distinguish perceptually between spontaneous and read aloud speech (e.g. [15, 16]), and [17] showed that there were less spectral distance between phonemes in spontaneous than in read aloud speech. In this section we will give an overview of the language composition of our spontaneous data and contrast it with the read aloud data. The composition of the spontaneous speech is not unique to our data, or to English, and similar distributions of fillers and grunts were reported for Japanese [13], and Spanish and Catalan [11] conversations.

2.2.1. Read Aloud Data

The read aloud data was recorded around the same time period as the conversational recording [10].

The sentences were recorded to provide phonetic coverage for speech synthesis, and consists of texts from a wide range of domains such as news, weather reports, addresses and also "conversation". The voice talent was requested to read them aloud in a natural but neutral manner. A total of 2717 sentences, approximately 100min of phonetic material (excl. silent pauses), were used for the voices in this paper.

2.3. Coverage and Composition

Table 1 gives a summary of the composition of the read aloud and spontaneous speech data. There were about 600 more utterances of read aloud data than spontaneous data. About a third of the spontaneous utterances consisted only of back-channels (e.g. *"yeah"*,*"okay"*,*"right"*) or short responses and confirmations (e.g. *"no I agree"* or *"that sucks"*). Although important for conversational interaction, their specific segmental and prosodic properties made their value for training and generating out-of-database propositional content questionable.

No. of	Conversational	Read Aloud
Utterances	2120	2717
Word tokens	19841	22363
Word types	2200	5026
Quinphone types	37654	58867

Table 1: *Overview of the content of the conversational and read aloud data.*

Among the most frequent words in both the conversational and read aloud data were short function words: *"the"*, *"a"*, *"you"*, *"I"*, *"of"*, *"to"*, etc. In table 2 these overlapping word types have been removed and show the remaining top five words from conversational and read aloud speech. The remaining top

five words of the read aloud speech contained rather arbitrary and meaningless words, but in the spontaneous speech the remaining top five contained words that are frequent because they are used to control turn-taking and express agreement or hesitation:

- *"yeah"*: as back-channel, turn initially or as filler
- *"know"*: part of the fillers *"you know"* and *"you know what I mean"*
- the filled pauses (*"um"* and *"uh"*), both as hesitation marker and turn regulating
- *"so"*: used frequently both turn/phrase initial and final

These phenomena were also virtually absent from the read aloud data with only three occurrences of *"yeah"*, two of the filler *"you know"* and no filled pauses. But whereas filled pauses and lexical fillers were very frequent in the conversational data, the "pure" grunts were more sparse (see table 3).

Conversational			Read Aloud		
rank	count	type	rank	count	type
1	818	yeah	10	204	he
8	344	know	12	192	one
10	318	uh	13	167	with
11	302	so	14	165	two
12	292	um	15	155	we

Table 2: *The 5 most frequent words and their rank in the conversational and read aloud data after removing overlapping word types.*

Another telling difference of the composition of the read aloud coverage material and the spontaneous speech was the trigram counts (including silent pauses). In the read aloud data 64 trigram types occurred five times or more, and only seven ten times or more, and no trigram occurred more than twenty times. In the spontaneous data 144 trigram types occurred more than five times, 82 occurred ten times or more, and 23 occurred twenty times or more. The majority of these frequent trigrams were either back-channels or around phrase boundaries, hence representing conventionalised means for starting, ending or keeping a turn.

2.3.1. Speaking Rate

Unit selection and HMM-based speech synthesis systems generally assume recordings of a consistent speaking style. Conversational speech however, has more variation and we will exemplify this with speaking rate.

The speaking rate of the conversational and read aloud data was measured for speech sequences delimited with silent pauses, measured as syllables per second. The variation of length of utterances was larger in the conversational data, and it is questionable if speaking rate is a relevant measure for back-channels, therefore the speech rate was only measured for utterances that were five to ten words long. Figure 1 show a boxplot of speaking rate. The conversational speech was on average eight percent faster than the read aloud speech, but more importantly there was much more variation in speaking rate between utterances for the conversational speech. This variation should be utilised better in HMM-based speech synthesis.

Conversational grunts						
oh	huh	ah	mhm	uh-huh	hmm	mm
34	18	6	6	5	4	4

Table 3: *Type and count of grunts in the conversational data.*

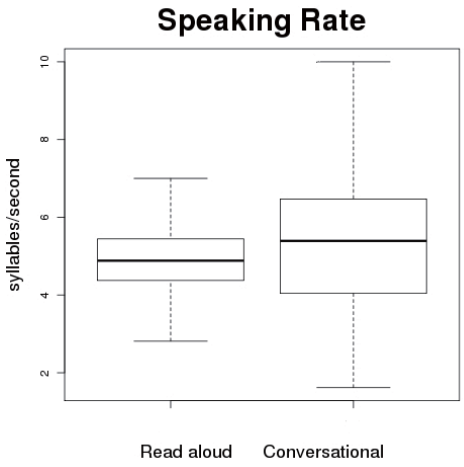


Figure 1: Speaking rate for utterances with 5-10 words in the conversational and read aloud data. The solid line is the median, box borders show the upper and lower quartiles, and the whiskers are drawn to 1.5 times the inter-quartile range.

2.4. Speech Processing

The automatic forced alignment, front-end linguistic analysis, and generation of context dependent phonemes for the HMM-based speech synthesis system of the read aloud sentences and the transcribed spontaneous speech were made with the CereVoice system [18].

Forced alignment from a flat start did not give sufficiently accurate phone alignment for the spontaneous speech. By utilising the trained acoustic models from the forced alignment of the read aloud speech on slowed down spontaneous speech, the forced alignment of spontaneous speech was improved, which gave a substantial improvement of synthetic speech quality. Utterance internal silent pauses was also detected and aligned in this step [10].

2.4.1. Linguistic Analysis

The CereVoice front-end provided a basic linguistic analysis well adapted for general read aloud text-to-speech use. But did not provide an accurate analysis of spontaneous speech phenomena like filled pauses, lexical fillers or conversational "grunts". However, accurate representations of spontaneous speech phenomena for HMM-based speech synthesis is an open research question, therefore we did not attempt a more sophisticated solution until we have assessed the extent and nature of the problem.

2.4.2. Context Dependent Phonemes

The context dependent phonemes defines the language related segmental and prosodic categories and dependencies in speech, for both the training and generation parts of HMM-based speech synthesis.

The contexts were based on [19] and were generated from the linguistic analysis and took into account segmental and prosodic contexts such as:

- quinphone (i.e. current phoneme and the two preceding and succeeding phonemes, example: s-p-o-r-t)
- position of phoneme in syllable, word and phrase
- position of syllable in word and phrase
- part-of-speech (content or function word)
- {preceding, current, succeeding} syllable stress (0/1) and accent (0/1)
- boundary tone of phrase (utterance final or medial)

Although the contexts at a glance seem a bit blunt, an important factor was that they often uniquely identified many of the spontaneous speech phenomena. E.g. *"yeah"* and *"um"* were the only two words with that unique combination of phonemes within a single syllable word. The quinphone context was also large enough to include both short function words (*"in"*, *"and"*, *"but"*, etc.) and common word endings like *"-ing"* together with a filled pause and thereby potentially preserving any associated hesitation in words immediately preceding a filled pause.

3. HMM-based Speech Synthesis

All the synthetic voices in this paper were built with the speaker dependent HMM-based speech synthesis system (HTS) described in [20]. The only difference between the voices were the speaking styles of the data and the additional blending of speaking styles (sec. 3.2). An overview of the acoustic feature extraction, training of HMM-based models, and generating synthetic speech in [20] is given below:

1. **Acoustic Feature Extraction:** Spectral and excitation parameters are extracted from the acoustic speech signal as STRAIGHT mel-cepstrals, aperiodicity and log F0.
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 HMM:s through decision tree-based context clustering. 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 and speech (spectral, excitation and duration) parameters are then generated from the corresponding trained HMM:s and rendered into a speech signal through the STRAIGHT mel-cepstral vocoder with mixed excitation.

3.1. Read Aloud and Spontaneous Voices

The context dependent phonemes generated with the CereVoice system (sec. 2.4.2) for the read aloud and spontaneous speech, was used to build one spontaneous and one read aloud synthetic voice with the HTS system (sec. 3).

The size of the clustered decision trees reflects the amount and complexity of the speech data. Table 4 shows that despite less data for the spontaneous than read aloud voice the clustered duration tree was larger for the spontaneous voice due to more variation. Unlike, for example, the mel-cepstral tree where the read aloud tree was larger due to more data and better phonetic coverage.

	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

Table 4: *Number of leaf nodes in the clustered duration, logF0, mel-cepstral and aperiodicity trees, for the spontaneous (SP) and read aloud (RD) voices. The ratio(SP/RD) shows the relative tree sizes.*

3.2. Blending Speaking Styles

To increase the phonetic coverage, and thereby improve general segmental and prosodic quality, while still preserving important conversational characteristics, the spontaneous and read aloud data were blended with a method previously used to blend and preserve different emotional speaking styles [4].

All the spontaneous and read aloud data were pooled in training, and an additional context: speaking style (spontaneous or read), were added to the context dependent phoneme descriptions. This context was then used during clustering to share mutual or sparse phonetic properties between spontaneous and read aloud speech, while avoiding to share frequent and distinguishing characteristics.

Detailed analysis of the sharing or splitting of spontaneous and read aloud speech properties remains to be done, but in the duration tree a split was made almost immediately based on the duration of the syllable nucleus. Whereas for excitation and spectral part the sharing or splitting seemed to be more complex.

During synthesis one of the speaking styles was selected by setting the speaking style context to either spontaneous or read aloud.

4. Evaluation

4.1. Natural or Conversational

The evaluation was designed to investigate two aspects of our synthetic voices: *"naturalness"* and *"conversational speaking style"*. The naturalness criteria have been extensively used for evaluating synthetic speech and gives information of overall speech quality, but evaluating a conversational or spontaneous speaking style have been less explored. In [10, 14] the questions about quality and style were asked together, but in this evaluation we wanted to investigate if speaking *style* could be evaluated separately from speech *quality*. Therefore the listeners were given only one of the questions: quality or style. To put focus on speaking style, rather than lexical content, the sentence

pairs in the listening test never contained the same utterances, e.g. *"yeah [pause] X-men is cool [pause] yeah"* was compared with *"right [pause] oh you have to to transcribe all this"*. For the style question the listeners were also explicitly requested to disregard the speech quality and try and focus on the style. The test was designed as a forced choice test where listeners had to express preference for one of the utterances in the pair. A total of fifteen utterance pairs were evaluated for each of the conditions in sections 4.2 and 4.3.

4.2. Spontaneous versus Read Aloud Speech Synthesis

The first comparison were between the synthetic voices built with either the spontaneous or the read aloud speech.

The test material were randomly selected from held-out transcripts of the spontaneous conversation. But with restrictions on the length and content of the selected material, so that the test sentences contained at least two filler items, and were between 5-15 words in total. Hence testing the voices ability to synthesise fillers and propositional content. Below are three of the fifteen selected test sentences:

- *"yeah [pause] X-men is cool [pause] yeah"*
- *"right [pause] oh you have to to transcribe all this"*
- *"so let's see [pause] but um [pause] yeah [pause] nothing exciting"*

4.3. Fillers versus No-Fillers

A pilot listening test indicated that blending read aloud and spontaneous speech resulted in better speech quality without loosing important conversational characteristics. To evaluate whether we could compete in naturalness with sentences without fillers and disfluencies and see how these phenomena affected the perception of speaking style, these phenomena were removed from the test sentences:

- *"X-men is cool."*
- *"You have to transcribe all this."*
- *"Let's see, but nothing exciting."*

The sentences without fillers were synthesised with the blended voice, with the speaking style context set to "read". And the sentences with fillers and disfluencies were synthesised with the blended voice, with the speaking style context set to "spontaneous". The comparisons in the listening test were then between e.g.: *"X-men is cool."* and *"right [pause] oh you have to to transcribe all this"*.

4.4. Listening Test

The listening test was carried out in a quiet lab environment and all listeners used headphones. Thirty two listeners, mainly native speakers of English, were payed to take part in the evaluation. Sixteen listeners were requested to evaluate *"naturalness"* and sixteen were requested to evaluate *"conversational style"*. The 15 sentences from the two conditions were randomised and mirrored, giving each listener 60 sentence pairs to evaluate.

5. Results

Figure 2 and 3 show the results for naturalness and conversational speaking style collapsed over all listeners. The results were calculated with the binomial test with two-sided 95% confidence interval and testing the null hypothesis that there were no preference between the voices in our comparisons.

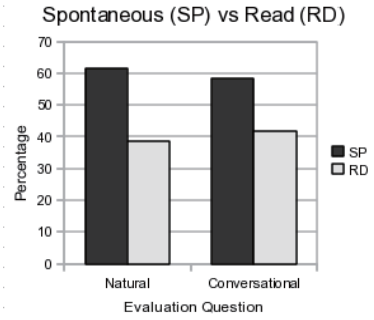


Figure 2: The bars show the percentages of the listeners' selections for naturalness and conversational style when comparing the spontaneous (SP) and read aloud (RD) voice when synthesising utterances with fillers.

The voice built with spontaneous speech was perceived as more natural (61.5% preference for spontaneous, 38.5% for read $p<0.05$) and had a more conversational speaking style (58.3% preference for spontaneous, 41.7% for read, $p<0.05$) compared to the voice built with read aloud speech.

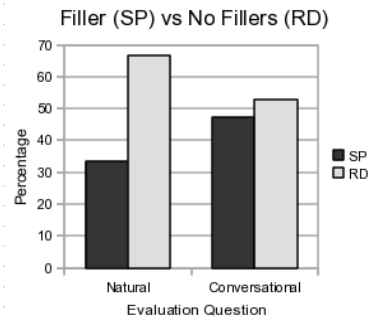


Figure 3: The bars show the percentages of the listeners' selections for naturalness and conversational style when comparing sentences with (SP) and without (RD) fillers synthesised with the blended voice.

The sentences without fillers were perceived as more natural than sentences with fillers (66.5% preference for without fillers, 33.5% preference for with fillers, $p<0.05$). But there were no significant difference in terms of conversational speaking style (52.7% preference for without fillers, 47.3% preference for with fillers, $p=0.25$).

6. Discussion

We already mentioned that duration and speaking rate in conversational speech is more complex than in consistently read aloud sentences and need to be better represented in HMM-based speech synthesis. But we have not mentioned the grunts (see table 3). Apart from *"oh"* the other grunts could not be synthesised with sufficient quality. Their sparsity together with their specific phonetic properties makes them unsuitable for sharing phonetic properties with other words, and currently the only solution would be to exclude them from training and use them as unmodified tokens.

This is in sharp contrast to the lexical fillers and filled pauses where the sheer amount was sufficient to build an HMM-

based synthetic voice with more natural sounding conversational characteristics than a voice built with carefully read aloud sentences, despite underspecified and erroneous linguistic analysis.

6.1. Evaluating Naturalness

The blending of spontaneous and read aloud speech did improve the general quality, but we could not synthesise speech with lexical fillers, filled pauses and disfluencies that could compete in "naturalness" with synthetic speech without these phenomena. Although we do need to improve analysis and representation of fillers and disfluencies for HMM-based speech synthesis, it was often the propositional content that sounded less natural, and not the fillers. The inclusion of disfluencies, although they do not stand out as unnatural, should probably have been avoided.

Comparing "naturalness" of isolated synthetic utterances of carefully articulated grammatical sentences and utterances with fillers and disfluencies might not be a relevant comparison. It is possible that listeners did not evaluate only "naturalness", but also took into account other aspects like: intelligibility, grammaticality or care of articulation.

6.2. Speaking Style

The evaluation of speaking style between synthetic voices built with either read aloud or spontaneous speech gave very similar results to the results for naturalness: the spontaneous voice had a more conversational style than the read aloud voice. This is not surprising given that the input text for both voices contained the same conversational phenomena, and therefore in the forced choice test listeners selected the voice with better quality.

Removing fillers and disfluencies from the test sentences did not result in a perceivable loss of conversational speaking style, and there were possibly several contributing factors to this result:

- Some listeners put more weight than others on speech quality in their decision about style.
- Other lexical items than the removed fillers contributed to a conversational style, e.g. *"...cool"*, *"...I could give a shit less..."* or *"...kind of a freak"*.
- The blending did, perhaps, not preserve distinct read aloud or spontaneous speaking styles. However such style blending might actually be attractive for virtual human characters to preserve important conversational characteristics, but with increased intelligibility over a voice built with only spontaneous speech.

7. Conclusions

We have demonstrated the importance of appropriate data for synthesising conversational characteristics with HMM-based speech synthesis. We have investigated blending of spontaneous and read aloud speech as a solution to the inherent problem of phonetic coverage in spontaneous speech data. But we have also highlighted potential difficulties of evaluating conversational speech quality and speaking style of isolated synthetic utterances.

8. Acknowledgements

We 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). This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk).

9. References

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *The Blizzard Challenge*, Edinburgh, U.K, 2009.

[2] D. Traum, W. Swartout, J. Gratch, and S. Marsella, "A virtual human dialogue model for non-team interaction," in *Recent Trends in Discourse and Dialogue*, L. Dybkjaer and W. Minker, Eds. Antwerp, Belgium: Springer, 2008, pp. 45–67.

[3] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer-Verlag Berlin Heidelberg, 2000.

[4] 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*, vol. E88-D, no. No. 3, 2005.

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

[6] N. Campbell, "On the structure of spoken language," in *Speech Prosody*, Dresden, Germany, 2006.

[7] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics & Cognition*, vol. 14, no. 1, pp. 129–182, 2006.

[8] J. Local and G. Walker, "Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech," *Phonetica*, vol. 62, pp. 120–130, 2005.

[9] E. Shriberg, "Phonetic consequences of speech disfluency," in *Proc. of International Congress of Phonetic Science*, San Fransisco, U.S.A., 1999, pp. 619–622.

[10] S. Andersson, K. Georgila, D. Traum, M. Aylett, and R. Clark, "Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection," in *Speech Prosody*, Chicago, USA, 2010.

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

[12] S. Sundaram and S. Narayanan, "Spoken language synthesis: Experiments in synthesis of spontaneous dialogues," in *IEEE Speech Synthesis Workshop*, Santa Monica, California, USA, 2002.

[13] N. Campbell, "Towards conversational speech synthesis; lessons learned from the expressive speech processing project," in *SSW6*, Bonn, Germany, 2007, pp. 22–27.

[14] 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.

[15] E. Blaauw, "Phonetic differences between read and spontaneous speech," in *ICSLP*, Banff, Canada, 1992, pp. 751–754.

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

[17] 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*, vol. 22, pp. pp.171–184, 2008.

[18] M. Aylett and C. Pidcock, "The CereVoice characterful speech synthesiser SDK," in *AISB'07*, Newcastle Upon Tyne, U.K., April 2007.

[19] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. of 2002 IEEE SSW*, 2002.

[20] 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*, vol. E90-D, no. 1, 2007.